Census 2000 Experiment

April 17, 2003

Administrative Records Experiment in 2000 (AREX 2000) Process Evaluation

FINAL REPORT

This research paper reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Michael A. Berning and
Ralph H. Cook

Planning, Research, and
Evaluation Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

Intentionally Blank

ACKNOWLEDGMENTS

The Administrative Records Experiment 2000 was conducted by the staff of Administrative Records Research at the U.S. Census Bureau, led by Charlene Leggieri. Questions and comments regarding this document can be directed to Michael A. Berning or Ralph Cook at 301-457-3067.

Administrative Records Research Staff Members and Key Contributors to AREX 2000:

Bashir Ahmed	Mikhail Batkhan	Mark Bauder
Mike Berning	Harold Bobbitt	Barry Bye
Benita Dawson	Joseph Conklin	Kathy Conklin
Gary Chappell	Ralph Cook	Ann Daniele
Matt Falkenstein	Eleni Franklin	James Farber
Mark Gorsak	Harley Heimovitz	Fred Holloman
David Hilnbrand	Dave Hubble	Robert Jeffrey
Dean Judson	Norman Kaplan	Vickie Kee
Francina Kerr	Jeong Kim	Myoung Ouk Kim
Charlene Leggieri	John Long	John Lukasiewicz
Mark Moran	Daniella Mungo	Esther Miller
Tamany Mulder	Nancy Osbourn	Arona Pistiner
Ron Prevost	Dean Resnick	Pamela Ricks
Paul Riley	Douglas Sater	Doug Scheffler
Kevin A. Shaw	Kevin M. Shaw	Larry Sink
Diane Simmons	Amy Symens-Smith	Cotty Smith
Herbert Thompson	Deborah Wagner	Phyllis Walton
Signe Wetrogan	David Word	Mary Untch

and

Members of the AREX 2000 Implementation Group

Intentionally Blank

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Intentionally Blank

EXECUTIVE SUMMARY

This report highlights the processes used for the Administrative Records Experiment 2000 and provides recommended improvements for future administrative records census operations.

The Administrative Records Experiment 2000 was part of the Census 2000 Testing, Experimentation, and Evaluation Program and was designed to gain information regarding the feasibility of conducting an administrative records census. An administrative records census is a census where housing and demographic data are drawn from administrative records from various government agencies. For the purpose of the Administrative Records Experiment 2000, records were drawn from the following agencies:

- Internal Revenue Service,
- Department of Housing and Urban Development,
- Center for Medicare and Medicaid Services Medicare,
- Indian Health Services, and
- Selective Service System,

The principal objectives of Administrative Records Experiment 2000 were to compare two methodologies for conducting an administrative records census to Census 2000 and to evaluate the results. Method 1 (referred to as the Top-down method) provides population counts down to the census block level. Method 2 (referred to as the Bottom-up method) attempts to match administrative records to the Master Address File and reconcile differences through field operations. This method provides both population and housing unit counts. Whereas both methods meet the data requirement for apportionment and redistricting, the Bottom-up method provides some additional data on housing unit relationship and tenure.

The experiment focused on five counties (two counties in Maryland and three counties in Colorado) that contained approximately one million housing units and a population of approximately two million persons. The sites were selected based on the mix of difficulty each represented in conducting an administrative records census. The operations for Administrative Records Experiment 2000 involved building a national database from the input source files and where appropriate, supplementing the record fields with data from other Census person and address records.

Basic results from the Administrative Records Experiment processing operations include:

- There is a reporting lag of approximately one year between the Statistical Administrative Records System 1999 /Administrative Records Experiment source files and the target date of April 1, 2000. The reporting lag impacted on our interpretation of results.
- Nationally, about 73 percent of Statistical Administrative Records System address records were machine geocoded. In Maryland, the machine geocoding rate was approximately 86 percent, while in Colorado the rate was approximately 80 percent.
- The clerical geocoding process added about three percent to the number of addresses geocoded in Maryland, and about five percent to the number of addresses geocoded in Colorado.

- For the Bottom-up method, administrative record addresses were computer matched to an April 2000 extract of the Decennial Master Address File. About 80 percent of Maryland Administrative Records Experiment addresses were computer matched to at least one Decennial Master Address File address, while about 81 percent of Colorado administrative record addresses were computer matched to at least one Decennial Master Address File address.
- A clerical review of the computer matching process added an additional four percent of addresses in Maryland and nearly six percent of addresses in Colorado by clerically matching addresses to the Decennial Master Address File.
- For administrative record addresses that did not match a Decennial Master Address File, field address verification was performed. The field verification was originally designed for 100 percent verification, but due to Census 2000 demand, the field verification was reduced to a sample basis composed of 6,644 addresses. About 13 percent of the Maryland addresses were valid as listed, while an additional 12 percent were deemed valid after the lister made minor corrections. In Colorado, about eight percent were valid as listed, and an additional 30 percent were deemed valid after minor corrections by the lister.
- The Administrative Record Experiment originally included a "Request for Physical Address" operation for addresses that were Post Office Boxes, commercial mailing services, and the like. This operation is evaluated in a separate report.

During the course of the experiment, several operations were modified from the original plan based on competing resources with decennial census operations. In spite of the changes, the Administrative Records Research Staff were able to adapt to the limitations and modify the operation to minimize the impact on the overall experiment. In lieu of a full-scale administrative records census, Administrative Record Experiment and Statistical Administrative Records System operations still may have many different applications to decennial census operations. An important example is imputation and Nonresponse Followup uses, which are discussed in the Administrative Records Experiment 2000 Household Evaluation. Such additional applications should be explored in 2000 – 2010 tests.

Time constraints did not allow for a detailed person-by-person comparison between the results of the Bottom-up method and the Decennial Census, nor between the results of the Bottom-up and Top-down methods. Although a household match was conducted between the Bottom-up method and the census, it remains an open question whether the matched addresses in the Bottom-up method contain the same people as those identified in the Decennial Census. Administrative Records Research should perform an evaluation using a detailed person-by-person comparison (micro-match) of the matched addresses within the Census and Bottom-up methods. Additionally, a detailed person-by-person comparison between the Bottom-up and Top-down methods should also be pursued with regard to person and address matches.

When the Administrative Record Experiment population tallies were produced and compared to the Census 2000 tallies, the results showed that for the Bottom-up method, the five test site county tallies, ranged from 96 percent to 102 percent of the Census 2000 population tallies. For the Top-down method, the range was 84-92 percent. Based on these results, we recommend

that administrative records continue to be tested and refined as a possible supplement for future census operations. Future refinement and improvements should, at a minimum, focus on the following areas:

- **Improve the computer matching and rematching processes**. An evaluation should be conducted to determine the effectiveness of the rematch to the Decennial Master Address File process. The dynamic nature of the Decennial Master Address File requires that it be continually updated from decennial census updates. Thus, duplicate and multiple Master Address File Identifiers for a given address may have changed since the first computer match. In addition, computer matching parameters must be further evaluated for accuracy and relevancy to the address matching task, as many addresses classified as possible matches by the computer were deemed to be matched during the clerical review process.
- **Evaluate the impact of multiple Master Address File Identifiers on the Decennial Master Address File**. Multiple Master Address File Identifiers assigned to a single address and duplicate Master Address File Identifiers assigned to multiple addresses contributed to the difficulty in classifying addresses as matched, non-matched, or possibly matched. Further research on the impact of retaining duplicate and multiple Master Address File Identifiers on the Decennial Master Address File should be pursued.
- **Improve the availability of source data for the under 18 population**. Administrative Records Research should continue to pursue coverage improvements via additional file acquisition. Expanding coverage of existing files should also be pursued in an attempt to improve coverage of certain segments of the population — particularly dependents on the Internal Revenue Service files and the under age 18 population segment nationally. Improving race information on administrative record files should also be pursued.
- **Evaluate the effectiveness of computer models used in the experiment**. Since the FAV Address Selection Model and the FAV Estimation Model influenced final tallies and results, further research should be conducted to assess the effectiveness of the models employed.
- **Conduct further research on address selection**. As the critical element for converting administrative record source data into a format useful for generating census tallies, a more thorough assessment of the StARS and Administrative Records Experiment address selection rules used to determine a person's "best address" should be pursued.
- **Conduct a full-scale field address verification**. Final Administrative Records Experiment results suggested an extremely limited ability to predict the number of valid addresses from a model. Using only a sample of addresses to conduct the field address verification operation, under the assumption that any addresses not matched to the Decennial Master Address File were true non-matches, led to the conclusion that only a full-scale field address verification operation would be acceptable.

Intentionally Blank

1. BACKGROUND

1.1 Introduction

The Administrative Records Experiment 2000 (AREX 2000) was an experiment in two areas of the country designed to gain information regarding the feasibility of conducting an administrative records census (ARC), or the use of administrative records in support of conventional decennial census processes. The first experiment of its kind, AREX 2000 was part of the Census 2000 Testing, Experimentation, and Evaluation Program. The focus of this program was to measure the effectiveness of new techniques, methodologies, and technologies for decennial census enumeration. The results of the testing lead to formulating recommendations for subsequent testing and ultimately to the design of the next decennial census.

Interest in taking a decennial census by administrative records dates back at least as far as a proposal by Alvey and Scheuren (1982) wherein records from the Internal Revenue Service (IRS) along with those of several other agencies might form the core of an administrative record census. Knott (1991) identified two basic ARC models: (1) the Top-down model that assembles administrative records from a number of sources, unduplicates them, assigns geographic codes and counts the results; and (2) the Bottom-up model that matches administrative records to a master address file, fills the addresses with individuals, resolves gaps and inconsistencies address by address, and counts the results. There have been a number of other calls for ARC research — see for example Myrskyla 1991; Myrskyla, Taeuber and Knott 1996; Czajka, Moreno and Shirm 1997; Bye 1997. All of the proposals fit either the Top-down or Bottom-up model described here.

Knott also suggested a composite Top-down/Bottom-up model, which would unduplicate administrative records using the Social Security Number (SSN) then match the address file and proceed as in the Bottom-up approach. In overall concept, AREX 2000 most closely resembles this composite approach.

More recently, direct use of administrative records in support of decennial applications was cited in several proposals during the Census 2000 debates on sampling for Nonresponse Followup (NRFU). The proposals ranged from direct substitution of administrative data for non-responding households (Zanutto, 1996; Zanutto and Zaslavsky, 1996; 1997; 2001), to augmenting the Master Address File development process with U.S. Postal Service address lists (Edmonston and Schultze, 1995:103). AREX 2000 provided the opportunity to explore the possibility of NRFU support.

The Administrative Records Research (ARR) staff of the Planning, Research, and Evaluation Division (PRED) performed the majority of coordination, design, file handling, and certain field operations of the experiment. Various other divisions within the Census Bureau, including Field Division, Decennial Systems and Contracts Management Office, Population Division, and Geography Division supported the ARR staff.

Throughout this report, rather than identifying individual workgroups or teams, we shall refer to the operational decisions made in support of AREX to be those of ARR; that is, we shall say that "ARR decided to…" whenever a key operational decision is described, even though, of course, ARR staff were not the only decision makers.

1.2 Administrative Record Census—Definition and Requirements

In the AREX, an administrative record census was defined as a process that relies primarily, but not necessarily exclusively, on administrative records to produce the population content of the decennial census short form with a strong focus on apportionment and redistricting requirements. Title 13, United States Code, directs the Census Bureau to provide state population counts to the President for the apportionment of Congressional seats within nine months of Census Day. In addition to total population counts by state, the decennial census must provide counts of the voting age population (18 and over) by race and Hispanic origin for small geographic areas, currently in the form of Census blocks, as prescribed by PL 94-171 (1975) and the Voting Rights Act (1964). These data are used to construct and evaluate state and local legislative districts.

Demographically, the AREX provided date of birth, race, Hispanic origin, and sex, although the latter is not required for apportionment or redistricting purposes. Geographically, the AREX operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in certain address matching operations and one of the evaluations; but generally, household and family composition were not captured. In addition, the design did not provide for the collection of sample long form population or housing data, needs that will presumably be met in the future by the American Community Survey program. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for the Census 2000.

1.3 AREX Objectives

The principal objectives of AREX 2000 were twofold. The first objective was to develop and compare two methods for conducting an administrative records census, one that used only administrative records and a second that added some conventional support to the process in order to complete the enumeration. The evaluation of the results also included a comparison to Census 2000 results in the experimental sites.

The second objective was to test the potential use of administrative records data for some part of the NRFU universe, or for the unclassified universe. Addresses that fall into the unclassified status have very limited information on them—so limited, in fact, that the address occupancy status must be imputed, and, conditional on being imputed "occupied", the entire household, including characteristics, must be imputed. In order to effectively use administrative records databases for substitution purposes; one must determine which kinds of administrative record households are most likely to yield similar demographic distributions to their corresponding census households.

Other more general objectives of the AREX included the collection of relevant information, available only in 2000, to support ongoing research and planning for administrative records use in the 2010 Census, and the comparison of an administrative records census to other potential 2010 methodologies. These evaluations and other data will provide assistance in planning major components of future decennial censuses, particularly those that have administrative records as their primary source of data.

1.4 AREX Top-down and Bottom-up Methods

1.4.1 Top-down

The AREX 2000 enumeration was accomplished by a two-phase process. The first phase involved the assembly and computer geocoding of records from a number of national administrative record systems, and unduplication of individuals within the combined systems. This was followed by two attempts to obtain and code physical addresses (clerical geocoding and request for physical address) for those that would not geocode by computer. Finally, there is a selection of "best" demographic characteristics for each individual and "best" street address within the experimental sites. Much of the computer processing for this phase was performed as part of the Statistical Administrative Records System (StARS) 1999 processing (Judson, 1999; Farber and Leggieri, 2002). As such, StARS 1999 was an integral part of AREX 2000 design.

One can think about the results of the Top-down process in two ways. First, counting the population at this point provides, in effect, an administrative-records-only census. That is, the enumeration includes only those individuals found in the administrative records, and there is no other support for the census outside of activities related to geocoding. AREX 2000 provides population counts from the Top-down phase so that the efficacy of an administrative-records-only census can be assessed.

However, without a national population register as its base, one might expect an enumeration that used only administrative records to be substantially incomplete. Therefore, a second way to think about the Top-down process is as a substitute for an initial mail-out in the context of a more conventional census that would include additional support for the enumeration.

1.4.2 Bottom-up

The fundamental difference between the Bottom-up method and the Top-down method is the Bottom-up method matches administrative records addresses to a separately developed "frame" of addresses, and based on this match, performs additional operations. In this experiment, an extract of the Census Bureau's Master Address File (MAF) served as the frame[1].

The second phase of the AREX 2000 design was an attempt to complete the administrative-records-only enumeration by the correction of errors in administrative records addresses through address verification (a coverage improvement analogue) and by adding persons missed in the administrative records (a NRFU analogue). This phase began by matching the addresses found in the Top-down process to the MAF in order to assess their validity and to identify those MAF addresses for which no administrative records were found. A field address review (FAV) was used to verify non-matched administrative records addresses, and invalid administrative records addresses were excluded from the Bottom-up selection of best address. Non-matched MAF addresses were canvassed in order to enumerate persons at addresses not found in the administrative records systems. In the AREX, such a canvassing was simulated by adding those persons found in the Census 2000 at the unmatched addresses to the adjusted administrative-

[1] In this report, we use the term "MAF" generically. Our operations were based on extracts from the Decennial Master Address File (DMAF).

records-only counts, thus completing the enumeration. Accomplishing the AREX as part of the Census 2000 obviated the need to mount a separate field operation to canvass unmatched MAF addresses.

Considering the Top-down and Bottom-up processes as part of one overall design, AREX can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a Top-down administrative records enumeration. Figure 1 below, provides a conceptual overview of the experiment for enumerating the population tested during the AREX. A more detailed description of data processing flows can be found in Attachment 1.

Note: The graphical description presented here is intended to convey the concept of both AREX methods when viewed in terms of the Bottom-up method as a follow-on process to the Top-down method.

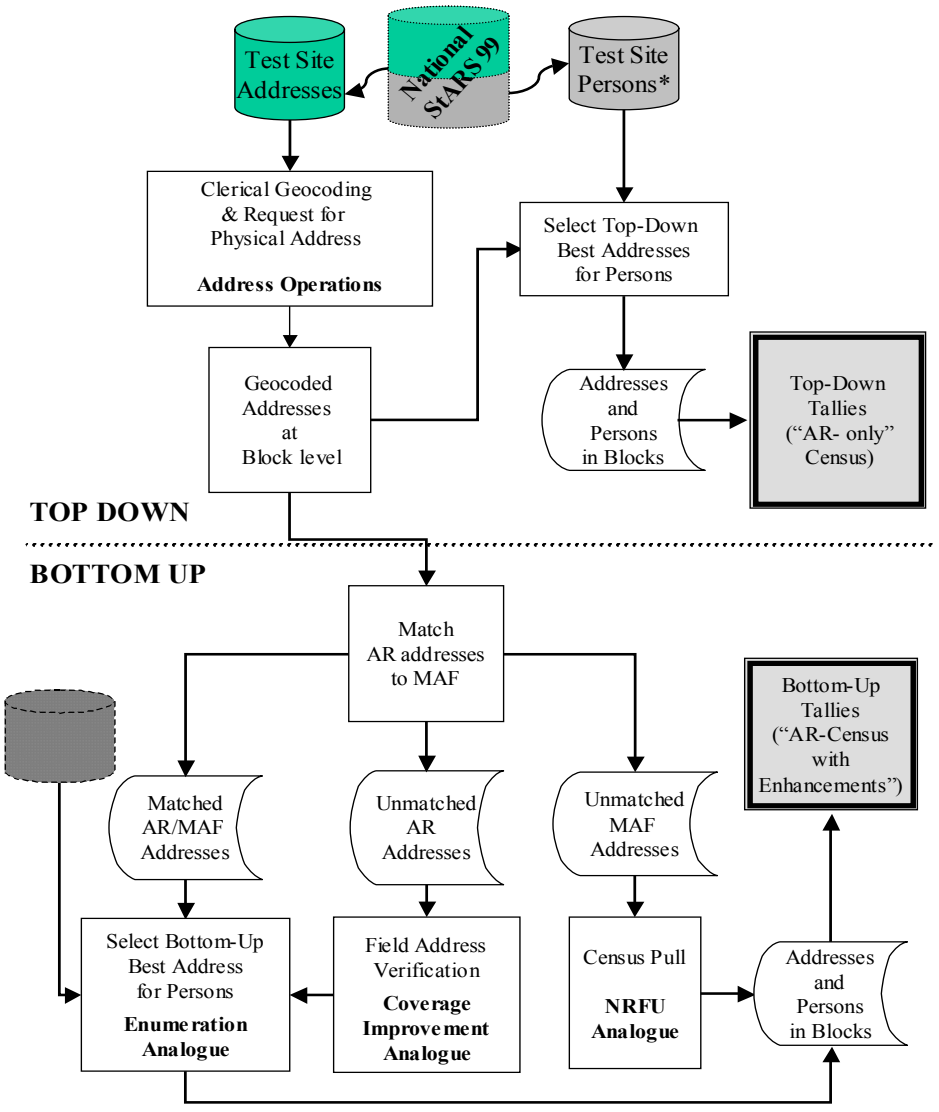


Figure 1. Summary Diagram of AREX 2000 Design

1.5 Experimental Sites

The experiment was set up to include geographic areas that include both difficult and easy to enumerate populations. Two sites were selected believed to have approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland. The other site included Douglas, El Paso, and Jefferson Counties, Colorado. The sites provided a mix of characteristics needed to assess the difficulties that might arise in conducting an administrative records census. Approximately one half of the test housing units was selected based on criteria assumed to be easy-to-capture in an administrative records census (for example, areas having a preponderance of city style addresses, single family housing units, older and less mobile populations), and the other half was selected based on criteria assumed to be hard to capture (the converse). Demographic characteristics of the sites are given in the following table.

Table 1. Key Demographic Characteristics of the AREX 2000 Sites

	Baltimore County, MD	Baltimore City, MD	Douglas County, CO	El Paso County, CO	Jefferson County, CO	United States
Total Population[1]	754,292	651,154	175,766	516,929	527,056	281,421,906
White[1]	74.4%	31.6%	92.8%	81.2%	90.6%	75.1%
Black[1]	20.1%	64.3%	1.0%	6.5%	0.9%	12.3%
American Indian, Eskimo or Aleut[1]	0.3%	0.3%	0.4%	0.9%	0.8%	0.9%
Asian or Pacific Islander[1]	3.2%	1.5%	2.6%	2.7%	2.4%	3.7%
Other Race[1]	0.6%	0.7%	1.4%	4.7%	3.2%	5.5%
Multi-Race[1]	1.4%	1.5%	1.9%	3.9%	2.2%	2.4%
Hispanic[1]	1.8%	1.7%	5.1%	11.3%	10.0%	12.5%
Median Age[1]	37.7 yrs	35.0 yrs	33.7 yrs	33.0 yrs	36.8 yrs	35.3 yrs
Crude Birth Rate[2]	12.6	14.9	19.0	15.7	12.5	14.9[3]
Crude Death Rate[2]	9.9	13.1	2.7	5.5	6.0	8.6[3]
1990-2000 Change[4]	9.0%	-11.5%	191.0%	30.2%	20.2%	13.2%

Note: all values include household and group quarters residents
[1] 2000 Census results
[2] 1998 rates per 1000; from MD Dept. of Health and Mental Hygiene and CO Dept. of Public Health and Environment
[3] 1998 rates per 1000; from www.fedstats.gov
[4] 1990 and 2000 Census results

1.6 AREX Source Files

The administrative records for AREX were drawn from the StARS 1999 database. There were six national-level source files selected for inclusion in StARS. Section 4.1 of this document describes the source files in detail. The files were chosen to provide the broadest coverage possible of the U.S. population, and to compensate for the weaknesses or lack of coverage of a given segment of the population inherent in any one source file. See Table 2 in Section 4.1.1 for a description of the source file characteristics.

1.6.1 Timing

An important limitation for the AREX is the gap between the reference period for data contained in each source file and the point-in-time reference of April 1, 2000 for the Census. The time lag has an impact on both population coverage—births, deaths, immigration and emigration—and geographic location—housing extant, and geographic mobility. As an example, both IRS files include data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999. Note, however, that the IRS 1040 file only provided persons in the tax unit as of December 31, 1998. The pertinent reference dates for each of the files is provided in Table 3, Section 4.1.1.

1.6.2 State, Local and Commercial Files

ARR staff decided not to use state and local files[2] and commercially available databases[3] in the AREX 2000 experiment. Statistical evidence is limited, but various reports from ARR staff indicated that state and local files come in an extremely diverse variety of forms, with equally diverse record layouts and content (for historical information, see Sweet, 1997; Buser, Huang, Kim, and Marquis, 1998; and other papers in the Administrative Records Memorandum Series). Furthermore, ARR staff reported that it was quite time-consuming and intricate to develop the interagency contractual arrangements necessary to use state and local files. Public opinion results such as Singer and Miller (1992), Aguirre International (1995), and Gellman (1997), convinced ARR staff that public sensitivity to the idea of linking commercial databases with government databases (other than for address processing) would be too great, and that such a linkage would be unwise.

In addition to acquisition and processing difficulties, consideration of the use of state and local files raises an equity issue in a decennial census context. Since it is not possible to obtain an exact count of the population in its entirety, public perception of fair treatment in the decennial census process is important. Therefore, the accuracy of the counts must be seen as uniform between and within states. The use of data from only certain states or localities would compromise notions that decennial census methods must treat all parts of the country equitably.

1.6.3 Census Numident

An additional, and critical, file used in creation of the StARS database was the Census Numident file. For the AREX, it was the source of most of the demographic characteristics and some of the death data. Detailed discussion regarding the creation and use of the Census Numident may also be found in Section 4.1.1.

1.7 AREX Evaluations

Currently, four evaluations are being completed.

The **Process Evaluation** documents and analyzes selected components or processes of the Top-down and Bottom-up methods in order to identify errors or deficiencies. It is designed to catalog

[2] Such as state and local tax returns, drivers license files, local utilities, assessor's records, and the like.

[3] Such as commercially available mailing lists, credit card databases, and the like.

the various processes by which raw administrative data became final AREX counts and attempts to identify the relative contributions of these various processes.

The **Outcomes Evaluation** is a comparison of Top-down and Bottom-up AREX counts by county, tract, and block level counts of the total population by race, Hispanic origin, age groups and gender, with comparable decennial census counts. This evaluation is outcome rather than process oriented.

The **Household Evaluation** assesses outcomes of the Bottom-up method, the potential for NRFU substitution and unclassified imputations, and predictive capability. NRFU substitution assesses the feasibility of using administrative records, in lieu of a field interview, to obtain data on non-responding census addresses via the Bottom-up method.

The **Request for Physical Address (RFPA) Evaluation** assesses the impact of noncity-style addresses. These addresses present a significant hurdle to the use of an administrative records census on either a supplemental or substitution basis is the determination of residential addresses and their associated geographic block level allocation for individuals whose administrative record address is a P.O. Box or Rural Route. AREX 2000 tested a possible solution in the form of the Request for Physical Address operation. Several thousand letters were mailed to P.O. Box and Rural Route addresses requesting the receiver to reply with their residential address for purposes of block level geocoding. This report documents in detail the planning and implementation of the operation. It also analyzes the results of the operation and assesses its potential future use as part of an ARC.

2. METHODOLOGY

2.1 General Questions

The primary goal of the Process Evaluation is to document the process used to transform the administrative records data in the counts that are assessed in later AREX evaluation reports. The general questions underlying all sections of the evaluation are:

1. What methods were used to create AREX counts?
2. How did those methods affect the final results of the experiment?
3. What recommendations should be made for new and improved administrative records processing as well as for future administrative records experiments?

2.2 Specific Questions and Methodology

Using project documentation and data files, analysis will focus on the following AREX 2000 components:

1. StARS Development and its relationship to the AREX.
2. Computer Matching to the Decennial Master Address File.
3. Clerical Review of the Computer Match Results.
4. Field Address Verification.
5. Producing tallies for the Bottom-up and Top-down methods.

2.3 Applying Quality Assurance Procedures

Quality assurance procedures were applied to the design, implementation, analysis, and preparation of this report. The procedures encompassed methodology, specification of project procedures and software, computer system design and review, development of clerical and computer procedures, and data analysis and report writing. A description of the procedures used is provided in the :Census 2000 Evaluation Program Quality Assurance Process."

3. LIMITS

This evaluation had three major limitations:

1. This evaluation will assume that there will be no changes in the content of national administrative record systems in the future that might facilitate their use in future decennial censuses. (For a discussion of some feasible changes, see Bye 1997.)
2. The evaluation will not address legal issues related to the use of administrative records for decennial census purposes or issues of public policy concerning the acceptability of these approaches by Congress or the public.
3. Concerning suggestions for additional administrative record sources that might be used in future experiments; the evaluation will be limited to data systems that are national in scope. It is assumed that the data processing difficulties associated with the use of State and local data are too great to warrant their consideration; and that the selection of data from just certain states or localities would compromise notions that decennial census methods must treat all parts of the country fairly and equally.

4. RESULTS

4.1 Building a National System of Administrative Records – StARS Development

4.1.1 Background and Overview

The prospect of conducting an administrative records census (ARC), where the sole source of data were administrative records, required development of a database on a national level. The Statistical Administrative Records System (StARS) is a research project designed to build a database of person and address data using administrative records from various government agencies, primarily for application to decennial census research and development. The StARS database was designed with two main goals required of the final output file:

> **person data:** One output record per person, assigned to an individual residence corresponding as closely as possible to Census residence definitions, containing characteristic data (age, gender, race and Hispanic origin), corresponding as closely as possible to Census short form data, and excluding persons which are not in the population of interest.

address data: One output record per address, geocoded to Census block with address and concepts corresponding as closely as possible to Decennial Master Address File (DMAF) address fields and concepts, and excluding locations which are not in the population of interest.

To this end, six national level source files were selected for inclusion in the StARS database. The multiple source files were selected based on the population universe associated with each file to provide the broadest coverage possible of the U.S. population. The national level files were selected to compensate for the weaknesses or lack of coverage of a given segment of the population inherent in any one source file. For the data content within each file, the Census Bureau requested the approximate content equivalent to Census "short form" data, or data that are used for other "long form" modeling (e.g., income) projects. In any case, programmatic data not immediately useful for decennial applications was not requested. At a minimum, each record on the file had to reflect a name, Social Security Number (SSN), and an address. Additionally, the source files had to be:

- releasable from the parent agency,
- transferable to a common medium to build a database,
- linkable to corresponding variables in a database, and
- evaluated for data quality.

The national level files that contributed to the StARS 1999 database, and therefore, to AREX 2000 included the following:

- Internal Revenue Service (IRS) Tax Year 1998 Individual Master File (1040),
- IRS Tax Year 1998 Information Returns File (W-2 / 1099),
- Department of Housing and Urban Development (HUD) 1999 Tenant Rental Assistance Certification System (TRACS) File,
- Center for Medicare and Medicaid Services (CMS) 1999 Medicare Enrollment Database (MEDB) File,
- Indian Health Services (IHS) 1999 Patient Registration System File, and
- Selective Service System (SSS) 1999 Registration File.

The following table displays the primary reason each file was included in the StARS database and the approximate number of input records associated with each:

Table 2. Source File Characteristics

File	Targeted Population Segment	Address Records	Person Records
IRS 1040	Taxpayer and other members of the tax reporting unit with current address	120 million	243 million
IRS W2/1099	Persons with taxable income who may not have filed tax returns	598 million	556 million
HUD TRACS	Low income housing population (possible non-taxpayers)	3 million	3 million
Medicare File	Elderly population (possible non-taxpayers)	57 million	57 million
IHS File	Native American population (possible non-taxpayers)	3 million	3 million
SSS File	Young male population (possible non-taxpayers)	14 million	13 million
	Total	**795 million**	**875 million**

Notes: Variance between the number of address records and person records within input source files is a result of the following file anomalies:

1. The number of address records column is generally synonymous with the total record count on the input file.
2. Each IRS 1040 input record may reflect up to six persons (primary filer, secondary, and four dependents).
3. Each SSS input record may reflect two addresses - defined as current and/or permanent address.
4. The IRS W-2/1099 file undergoes a preliminary unduplication and clean-up prior to the initial file edit process. Prior to person processing, records are written "out of scope" if the SSN field is blank, the name standardizer returns a "bad name", or the edited or input name field is blank. "Bad names" include institutional or firm names not recognized by the standardizer.

To achieve the desired person and address data output goals, a "dual-stream" processing approach was adopted to derive the best possible geographic and demographic data for each record. During the initial file edit phase a unique record identifier (UID) was assigned to each source input record, and separate files for person and address data were created from each of the source files to enable dual stream processing. The UID assignment to each record ensured the capability to re-link the person record with the correct address record. A series of pointer files were created and updated along the various processing steps to enhance the re-link capability. Dual stream processing also provided the capability to de-conflict the geographic and demographic data within each source file **and** among the various source files in a more objective fashion. The six source files were edited to standardize name, date fields, and other demographic information, then combined and passed through a Social Security Number (SSN) validation algorithm. The validation process compared the source record SSNs against the Census Numident file, and where appropriate, "filled" any missing demographic characteristics on the source record with the data present on the Census Numident. The resulting verified person records with complete demographic characteristics were then ready for re-linking to addresses.

Address information from the source files also underwent editing, standardization, unduplication, and "best data selection" processes, and then re-linked to the verified person records. Processing continued with application of selection criteria rules for each of the demographic characteristics and selection of the best address. The resulting output was the Composite Person Record (CPR) — for all intents and purposes, the StARS database.

A critical limiting factor to AREX was the time lag between the data contained in each source file and the "moment in time" comparison against Census 2000 results. As an example, both IRS files included data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999. Although the file cut date was requested for as close to April 1, 1999 as possible, the nature of IRS processing (assuming filing extensions, amended returns, etc.)

dictated for StARS 1999 that only processing cycle weeks 1—39 for the IRS 1040, and cycle weeks 1—41 for the W-2 / 1099 file be incorporated into the StARS database. The remaining processing cycle weeks files did not arrive at Census until the following February. Generally, the reference point for the address information was April 1, 1999. It should be noted that the IRS 1040 file provided persons in the tax-filing unit (tax return) as of December 31, 1998.

In any event, the latest updates to any of the six source files would be at least one year prior to the Census 2000 date. The following table provides a display of each file's relative currency to the April 1 Census 2000 date.

Table 3. Currency of Source Files

Source File	Cut-off Date	Requested Cut Date	Universe
Indian Health Svc.	04/01/99	04/01/99	All persons alive at cut-off date
Selective Service	Note 2	04/01/99	Males between the age of 18 - 25[2]
HUD TRACS	04/01/99	04/01/99	All persons on file as of cut-off date
Medicare	Note 3	04/01/99	All persons alive at cut-off date
IRS 1040	12/98	09/30/99[1]	Individual tax returns for tax year 1998
IRS W-2 / 1099	12/98	04/01/99[1]	Forms W-2 and 1099 forms for tax year 1998

1. File Cut date is for posting cycle weeks 1-39 only for IRS 1040, and weeks 1-41 for IRS 1099 files. Weeks 40-52 (and 42-52 respectively) were not included in StARS '99.
2. Cut-off date is same as dates used to define universe: persons born after April 2, 1972 and on (or before) April 1, 1980.
3. Universe also defined as persons with a death date of 12/31/1989 or later.

The IRS 1040 file was viewed as the primary source file to the StARS database with the greatest likelihood for reflecting the most current address for a given housing unit (the belief that persons expecting to receive a tax refund check will provide accurate address information). The remaining five source files were selected for inclusion in StARS for specific segments of the population that may not routinely file an annual tax return, or may supplement or amplify data reflected in the 1040 file.

An additional, and critical, file used in creation of the StARS database was the Census Numident file. The Census Numident was created by ARR for the primary purpose of validating Social Security Numbers (SSNs) used in the processing of administrative records and supplying demographic variables missing from source files. The Census Numident is an edited version of the Social Security Administration's (SSA) Numerical Identification (Numident) File. The SSA Numident file is the numerically ordered master file of assigned Social Security Numbers that may contain up to 300 entries for each SSN record, although on average contains two records per SSN. Each entry represents an initial application for an SSN or an addition or change (referred to as a transaction) to the information pertaining to a given SSN. The SSA Numident contains ***all*** transactions (and therefore, multiple entries) ever recorded against a single SSN. The SSA Numident available for StARS 1999 reflected all transactions through December 1998.

The Census Numident was designed to collapse the SSA Numident entries to reflect "one best record" for each SSN containing the "best" demographic data for each SSN on the file. However, all variations in name data (including married names, maiden names, nicknames, etc.)

and all variations of date of birth data were retained as part of the Census Numident as an Alternate Name File and Alternate Date of Birth File, respectively. Selection criteria were established for each of the desired Census 2000 Short Form demographic variables (after minor edits were accomplished in an effort to standardize the variables). The short form variables included such items as date of birth fields, gender, race, and Hispanic origin. Following the editing, unduplication, and selection processing, the SSA Numident file of nearly 677 million records was reduced to just over 396 million records within the Census Numident file.

Still another file created by PRED and used in StARS processing was the Person Characteristics File (PCF). During creation of the StARS CPR, ARR staff overcame a recognized limitation of the source files by producing PCF, which incorporated modeling techniques to impute a probability of race, gender, and mortality. The race model was designed to smooth out the inconsistent reporting and quality of race and ethnicity data present in the administrative records. The Numident file, which provided the widest coverage of race and ethnicity, in particular contained many race inconsistencies. Prior to 1980, SSA limited the race categories to "Black," "White" or "Other", while census definitions at the time included Asian or Pacific Islander (API) and American Indian. In the early 1980s, SSA expanded the race categories (for new applicants) to include API, Hispanic, and American Indian or Eskimo. However, race reporting is voluntary on the SSA application and therefore not always present on the file. Since 1986, SSA does not record the race or ethnicity for infants assigned an SSN at birth. Census 2000 expanded the race categories even further by separating Hawaiian and Pacific Islander from Asian and allowing all respondents to self-report multiple race categories. The StARS administrative source records have yet to adopt the expanded features of race reporting. Thus, a race model was created to impute a single best race based on the probability distribution produced by the model.

The gender model, based on the strength of association between first and middle names and reported sex, imputed a sex when the data were missing from the person record. A mortality model, using historical mortality rates computed by national health agencies and other factors - such as birth date, created a probability of current mortality for every person in the Census Numident. The PCF was the product of applying the three models to the Census Numident. Whenever demographic information was missing for a person record from one of the StARS source files, the PCF imputed (modeled) value was placed on the CPR.

4.1.2 StARS Processing

An overview of the "dual-stream" processing methodology employed during creation of the StARS database is provided in the following paragraphs. A more detailed description (in outline format) of StARS processing may be found in Attachment 2 to this document. Processing was accomplished in six main phases that included:

- File edit programs for address data,
- Code-1 address processing and geocoding the address records,
- Creation of a Master Housing File,
- File edit programs for person data,
- SSN Verification of person records, and

- Unduplication of person records, and creation of the Composite Person Record — the final output of the StARS database, which presents the "best" address and demographic characteristics for each person record.

For AREX 2000, test site addresses were extracted from the geocoded files and processed through AREX specific operations. The re-link of AREX address records to the appropriate person records did not occur until the AREX post-processing operations, which produced the baseline tallies for both AREX methods.

4.1.3 StARS Address Data Processing

Address data from the raw input files were formatted to meet StARS database requirements. The initial file edits were designed to prepare the address records for processing through a Group-1 Software Incorporated product known as Code-1. The Code-1 software product matched a file of address records against a national database of mailing addresses [certified and corresponding to United States Postal Service (USPS) standards]. Where Code-1 matched an input address, the address was standardized and updated to match the USPS reference file. If the input address was not matched, the Code-1 product standardized the input address. ARR staff viewed the Code-1 process as "cleaning" the address data prior to forwarding the records to the Geography Division (GEO) for input to the geocoding operation[4]. The geocoding operation was designed to further match addresses against the Topologically Integrated Geographic Encoding and Referencing (TIGER) database program.

The identification and disposition of "proxy" address[5] data within the source files was a particular processing issue. Proxy addresses were identified with flags to ensure proper consideration under the address selection rules during creation of the CPR. Preliminary research into the proxy address features indicated that no definitive conclusions could be drawn regarding identification of the correct "owner" of the proxy address. Therefore, the address selection rules invoked during CPR creation preferred selecting a non-proxy address over an address identified with a proxy flag.

Prior to the Code-1 operation, the address records were split into 1,000 cuts (000 - 999) based on the first three digits of the ZIP Code. The first cut (000) was reserved for those addresses reflecting invalid (non-numeric) or blank ZIP Codes. This first cut was not forwarded to GEO for the geocoding operation. Once the entire array of addresses were processed through Code-1, records were reassembled to place all addresses in the correct 3-digit ZIP Code cut. Within each cut, the address records were unduplicated based on an exact match of the street addresses and the full 9-digit ZIP Code[6] in an effort to reduce the number of address records forwarded to the Geography Division for the geocoding process. Following the unduplication effort, the 999 cuts, consisting of nearly 148 million records, were forwarded to GEO for the geocoding operation.

---

[4]The assignment of an address, structure, key geographic location, or business name to a location that is identified by one or more geographic codes.

[5]A proxy address is defined as a person or institution that receives mail on behalf of another individual - in this case the record holder. Proxy addresses may be identified by such terms as " in care of, "for," and "c/o". A separate file was created to house the record identifier and the proxy flag.

[6]See section 5.1.10 for a discussion regarding the impact of unduplication on the **full** 9-digit ZIP code (ZIP + 4).

During geocoding of the StARS addresses, potential addresses for inclusion in AREX 2000 test sites were identified by the Geography Division based on ZIP Code. As addresses were passed through the geocoding system, if the system placed an address in a collection block recognized by TIGER, the address was flagged as located within the test site. The machine geocoding process was designed to return address records to ARR with Census 2000 collection block geography and TIGERLINE ID numbers. The collection block and TIGERLINE ID data were essential to downstream processes in AREX 2000. As can be seen in the table, the geocoding rates for both Maryland and Colorado test site counties exceeded the national average.

Table 4. National Geocoding Tallies

	# Input Records to Geocoding	# of Records Geocoded	Percent Geocoded
StARS National Address File	147,346,145	108,032,169	73%
Maryland subset of StARS National File	725,108	626,247	86%
Colorado subset of StARS National File	624,248	498,783	80%

4.1.4 Creating the Master Housing File

Once the geocoded files were returned, ARR staff commenced creation of the Master Housing File (MHF). Building the MHF required an attempted match of records in the geocoded file to a file of commercial addresses maintained by American Business Information (ABI), Inc. Since known commercial addresses were not to be included in the AREX 2000 population tallies, all such addresses in the geocoded file were identified with a "commercial flag" for further AREX and StARS processes. The MHF was created from a series of files and processes that employed many temporary files, processing actions, and complex decisions. A summary of the basic steps to create the MHF follows:

First, the ABI file of commercial addresses was processed through Code-1 to standardize commercial address fields in the same format as that used to process the six national level files. Here, the ABI file was also split into 1,000 3-digit ZIP code cuts to facilitate a merge with the geocoded file. Once processed through Code-1, the ABI file *and* the geocoded file addresses were passed through a version of GEO's address standardizer. This additional standardization of the address fields assisted in the final unduplication of records in an attempt to display only unique records in the MHF. The return of parsed address fields by the standardizer permitted categorization of addresses by twelve types, which was critical in the construction of a Housing Unit Identifier (HUID). The HUID, a numeric identifier that replaced the many possible variations of an input street address, was the key element used in selection of a "best address" for retention on the Composite Person Record.

Next, the records were unduplicated a final time and matched against the ABI file to identify and flag commercial address records. Unduplication of the records was accomplished using a complex unduplication key that checked for the presence of TIGERLINE ID, state codes, ZIP Codes, and certain address standardizer return fields. A single record from among the geocoded duplicates was selected for retention on the MHF after a comparison check of Code-1

intelligence flags and GEO confidence flags that distinguished the degree of reliability in the Code-1 or GEO matching process. Non-geocoded records were also unduplicated in the same fashion using all but the TIGERLINE ID. In addition, certain "bad address" types (blank, non-parsed, and undefined addresses) were, essentially, unduplicated within each 3-digit ZIP Code as only one representative address for each of these types was retained on the MHF. A Master Pointer File was updated to reflect all duplicate records that existed within the database, which were represented by the unique address retained on the MHF. Once the MHF was created, each address record was ready for re-linking to its corresponding person record.

Where an exact match of the ABI file and the geocoded file occurred, the record was assigned a commercial flag and certain commercial variable fields were moved to and retained on the MHF. It must be noted that addresses were matched to the ABI file at a "basic street address (BSA) " level only, to flag such records as "potential" business/commercial addresses. The BSA included only house number and street name information. Designations for apartments, units, or lots were not included in the BSA. In effect, all units at the BSA received a commercial flag. As an example: a four-unit apartment structure at 101 Main Street may include one unit as a real estate office that ABI recognized as a commercial address. The remaining three residential apartments at 101 Main Street would also be assigned the commercial flag. Such records would ***not*** be unduplicated; rather all units at this particular BSA would receive the commercial flag. The difficulty in selection of a "best" address in these situations was readily apparent. In StARS '99, addresses with a commercial flag, were selected at a lower priority than non-commercial addresses. The "residential" units at the example cited above were less likely to be selected as a "best address" in StARS — all other selection criteria being equal.

4.1.5 StARS Person Data Processing

The StARS person edit process standardized and parsed names from the six source files and recoded common demographic variables to conform to Census Numident format. The records then underwent an SSN verification process against the Census Numident. Records were verified based on matching criteria for the SSN, name data, and date of birth (non-IRS records only) data. Demographic data from the Census Numident were appended to the verified IRS records. Any missing demographic data for non-IRS records were also appended from the Census Numident to the verified records. Records not initially verified, underwent a further search process using additional matching criteria within a commercial software matching program known as AutoMatch. Records not matched in the search process were retained in a separate unverified file of SSNs, and not processed further or included in the final StARS CPR. Correct address data were re-linked to all records in a later StARS process. The SSN verification rates within the StARS database for each source file are displayed in the following table:

Table 5. SSN (Person Record) Verification Results

Source File	# Person Records	Records to Verification	# Records Verified	# Records to Search[1]	# Found in Search	Total Valid Records[3] % of Total	Total Invalid
IRS 1040	243,260,776	243,260,776	238,309,801	1,783,628	214,834	238,562,104 97.9%	4,698,672[2]
W-2 / 1099	556,039,480	556,039,480	530,786,604	25,252,876	113,423	530,910,525 95.5%	25,128,955
MEDB	56,836,356	56,593,743	56,361,257	475,099	196,216	56,557,850 99.6%	278,506
HIS	3,095,928	2,526,201	2,441,761	654,144	193,434	2,635,945 96.7%	459,960[2]
SSS	13,176,234	13,063,105	12,681,966	494,268	370,491	13,055,125 97.1%	121,109
HUD TRACS	3,342,199	3,232,389	3,022,628	319,546	197,688	3,223,747 93.5%	118,427[2]
Total	875,750,973	874,715,694	843,604,017	28,979,561	1,286,086	844,945,296 96.4%	30,805,677[2]

1. Total records to search includes records from non-IRS files where no SSN is present on the record.
2. IRS 1040 records not passed to search (due to lack of full name, date of birth, and gender) totaled 3,167,347. Of these 3,160,445 were dependents not eligible for search due to the lack of name data. Only 48 additional records were not eligible for search from other source files – 25 HUD TRACS and 23 IHS records. The "Total Invalid" column includes the records from these three source files.
3. The total valid column also includes 55,193 records deemed valid - flagged code 9, which indicates the SSN appeared only on a quarterly update to the Census Numident.

4.1.6 Creating the Person Characteristics File (PCF)

The PCF modeled a race, gender, and mortality status for every person record present in the Census Numident regardless if the demographic data were present on a given Numident record. Modeling requirements were established early in the PCF development process to ensure that each model would provide a non-blank value for every output field and that an output record would be created for every input (Census Numident) record.

Race model data were constructed, primarily, based on last name and gender (place of birth was also a factor). The Census Bureau's most current Asian and Spanish surname lists were used in the race model construction. Additionally, the American Indian indicator field from the Indian Health Services source file was incorporated into the race model. The race model also determined the probability of Hispanic origin values output to the PCF.

The mortality model was constructed from a Death/Survival database created by ARR specifically for use in construction of the PCF. Age calculations, based on a cut-off date of April 1, 2000 (Census Day), were incorporated into the model, and all persons with a calculated age greater than 119 were assumed to be deceased. The database was drawn from the following sources:

- Date of birth information from the SSA Numident file.
- Reported date of death from the Medicare and IRS 1040 files.
- Date of birth information from the Medicare, Selective Service, and IHS files.

Gender model data were based on first name data present on the input record. Look-up tables containing common names, uncommon names, name-gender proportions, and gender model parameters were created and a final gender probability assigned after the four look-up tables were created and run against each input record. Each of the models were output in 20 segments split by SSN in the same fashion as the Census Numident. The resulting PCF was also output in the 20-segment format to facilitate merging with the verified records from the StARS SSN Verification process.

The "best data selection" rules alone were not expected to resolve all such demographic conflicts during StARS person processing. To account for this anomaly, PRED established a requirement to generate a modeled race on the PCF for every record present in the Census Numident. Since the PCF was created for uses other than input to the StARS database, the modeled demographic values and reported demographic values (if any) for a given person record were both output on the PCF. However, once all records were unduplicated during StARS person processing, a person's modeled race, gender, or mortality status was selected from the PCF only in the case where no race, gender, or mortality appeared on any administrative record for that given person record**.** Additionally, the PCF modeled data were used as a tiebreaker in certain cases where conflict appeared among the source records during selection of the "best" demographic data.

4.1.7 Creating the Linked Person Records

Before the Composite Person Record (CPR) was created, the re-linking of correct address data and person data was required. The purpose of the CPR was to present the "best" demographic and geographic data associated with each verified SSN in the StARS database. The intermediate process of linking person records provided a means to view all data for each SSN record, and then selection of the "best" data from among all possible values — thus creating a "composite" record for a given SSN.

At this point in the StARS database creation, the "dual stream" processing of address and person data were brought together in preparation for creating the Composite Person Record. The best address data from the MHF were merged with the person records that underwent the SSN verification process and the application of modeled demographic data from the PCF. Note that no address or demographic data were removed from the database (i.e., duplicate address records are identified via a series of pointer files). Likewise, no person records were removed from the database. It is known that duplicate addresses (as well as multiple addresses) existed for a given SSN record. The hypothetical records in Figure 2 (below) help to illustrate the unduplication methodology employed in the "dual-stream" process.

During the first unduplication of addresses, prior to the geocoding process by GEO Division, records 1 and 2 (in the example figure) were not unduplicated due to a variance in the street type (St. versus Rd.) and the full 9-digit ZIP Code. Once the geocoded records were returned, records were unduplicated a second time prior to creation of the Master Housing File. In this case, since the TIGERIDs were identical for records 1 and 2, the unduplication key identified and selected record 1 for retention on the MHF (geocoding "confidence" flags and Code-1 "intelligence" flags are also employed in the unduplication process). An HUID was constructed for record 1 and the Master Pointer File was updated to indicate a duplicate address record existed for this "selected" address. Records 3 and 4 would have been unduplicated prior to the geocoding operation, and upon return of the geocoded file, only one HUID assigned to this "unique" address. The master pointer file contained all data identifiers (address identifiers [AIDs] and unique record identifiers [UIDs]) to enable linkage of person records with appropriate address records. For the purposes of address processing, it was irrelevant that administrative records reflected "Thomas Jones" with three addresses, or that one of his addresses was also reflected on a different person record. The application of address selection rules during creation of the CPR ultimately selected only one of the three possible addresses for Thomas Jones and, by default, selected address record 4 for George Smith.

	Name	Address			ZIP Code	TIGERID
1.	Thomas Jones	127 Oak St	Non Site	State	62886 2258	9283710661
2.	Thomas Jones	127 Oak Rd	Non Site	State	62886 0000	9283710661
3.	Thomas Jones	1246 Sutton St	In Site	MD	21852 2357	1088653227
4.	George Smith	1246 Sutton St	In Site	MD	21852 2357	1088653227

Figure 2. Record Unduplication Example

During the SSN Verification process, all four records in the example were passed through the verification (and search if required) process. That is, no unduplication of records based on SSN or name data was accomplished prior to, during, or after the SSN Search and Verification process. For our example, assume that all three records for "Thomas Jones" reflected an identical (and verified) SSN. Each "Thomas Jones" record was carried forward to the Linked Person File where the correct address was placed on the person record based on what the master pointer file dictated. In this case, records 1 and 2 were assigned identical HUIDs as were records 3 and 4. Only one record for Thomas Jones could be output to the CPR. Thus, the unduplication of person records from the "dual stream" process was accomplished during application of the address and demographic selection rules. The rules were applied against all records in the Linked Person File. In the example, the CPR output would most likely reflect Record 1 for Thomas Jones and Record 4 for George Smith.

It must be noted that creation of the AREX Person Universe File prior to the post-processing phase of the AREX experiment, required all four records from the example to be included in the AREX person universe. Since AREX is a subset of the StARS database, the AREX person records resulting from the re-link of person and address records reflected a single StARS "best" address based on selection rules geared toward a national database. To ensure appropriate address records were available for determining a person's eligibility for inclusion in the AREX test site tallies, all address records for any person ***ever*** associated with an AREX test site address were placed in the AREX Person Universe File. The address selection rules were re-applied to

the AREX person universe to identify persons "in" or "out of scope" for AREX tally purposes. Looking at our example, Record 3 for Thomas Jones would be flagged "out of scope" since his best address was identified as Record 1. George Smith on Record 4 would be included in the final AREX tally file. The example provides an indication that Thomas Jones was an "out mover" from the AREX test site.

Two technical processing issues complicated the re-link process: 1) returning proxy address data to the address selection equation, and 2) establishing a methodology to deal with the multiple addresses that may have appeared within a single input record on the Selective Service File. During the initial address file edit process, an indicator flag to define proxy address data was stripped from each record and placed on a separate file - ostensibly to facilitate research on proxy addresses[7] During selection of a "best" address for a given SSN record, proxy address information was included in the address selection criteria - thus the requirement to re-link the proxy flag file with the correct person record. The Selective Service multiple address issue was resolved by using only the "current" address in downstream processing, since the "current" address was considered the better address for simulating a census enumeration address.

The primary steps required to re-link the address and person records follow:

1. The Master Housing and Master Pointer files were merged by 3-digit ZIP code.
2. Specific geography variables were extracted from the merged file to create a temporary Enhanced Master Pointer File (EMPF), which was split back to original input source file cut and sequence order. During the re-split, only the current Selective Service System address was retained on the EMPF for consideration during the address selection process.
3. Likewise, from the Selective Service System Proxy file, only the proxy flag data for a "current" address were selected for retention on the EMPF.
4. The EMPF and Proxy files were merged to create an address file that contained HUIDs, other geographic variables desired for the CPR, and proxy address data.
5. A direct access method was employed to link the person records with the correct EMPF records. The verified person records were used as the "driver" wherein each SSN record is read-in. The SSN unique record identifier was analyzed to determine which EMPF source file contained the matching unique record identifier. Once the correct record was found, the geographic data from the EMPF were appended to the SSN "driver" file record. In this fashion, all SSN records (including the unverified records) were linked to the correct address data.

4.1.8 Creating the Composite Person Record (CPR)

At this point in the process, duplicate person records existed in the StARS database. The process of selecting the best demographic and geographic data from among the linked person records resulted in the CPR output file. Where duplicate, linked person records were present, the "best" data were selected from among the entire array of any duplicate records. Thus, a true "composite person" record was retained in the StARS database. Creation of the Composite Person Record

[7] Later analysis revealed this as an unnecessary step in StARS processing. Retention of the proxy flag (a one character field) on the edited address files throughout the process was adopted for StARS 2000 production.

represented a final unduplication of person records from within the entire StARS database. The processing methodology employed to create the CPR follows:

1. Linked Person Files were in SSN sort order to allow for "SSN by group" processing. A processing array was established for each of the demographic variables and the address (HUID).
2. The Person Characteristics File (PCF) was opened and imputation probabilities and a special SSN encryption key known as a PIK (Protected Identification Key) were appended to the Linked Person Files. The PCF contained the Census Numident demographic values as well as imputed values for the demographics based on modeled data. Where the linked person record reflected blank demographic values, a PCF modeled value was output to the CPR. Thus, each unique record that appeared on the CPR reflected a race, gender, Hispanic origin, and PCF modeled probability for date of death.

Selection rules were invoked for the address (HUID) as well as each of the demographic variables. A more thorough discussion of the variable selection rules may be found in Attachment 2. Generally, a highest score or frequency of observation was the primary selection rule. In the case of address selection, geocoded addresses were selected over non-geocoded addresses. The decision to prioritize address selection by geocode success rate was made with the goal of supporting AREX 2000 requirements — the conduct of an administrative records census paralleling decennial census operations as closely as possible. As a starting point for the AREX Bottom-up method, an address list reflecting geocoded addresses simulated use of the Decennial Master Address File during decennial census operations.

4.1.9 StARS Extracts for AREX 2000

In simple terms, AREX 2000 address and person data may be viewed as subsets of the StARS database. In reality, the two data sets were treated differently as an initial AREX Address File (AAF) was created from the StARS address database when Geography Division flagged addresses within the AREX test sites based on ZIP Code and TIGER database information. The AAF underwent several iterations and operations to refine the data within the AAF until the AREX Post-Processing phase of operations dictated the requirement to re-link AAF data with StARS Composite Person Records.

From a national-level standpoint, StARS was primarily concerned with unduplicating records to ensure counting a person only once. The linking of address and person records for the AREX test site subset of StARS had to also consider the point in time at which an AREX address was still valid for a given person record. Returning to the hypothetical example in Figure 1, only one of the persons could be placed at the address for records 3 and 4. AREX post-processing had to account for the "in/out" mover — either Thomas Jones or George Smith. In other words, was the AREX address the "best address" for each person identified as residing within the AREX test site? To answer the question, a universe of AREX addresses and AREX persons was created at different points in time with regard to the StARS database.

The AREX address universe was born out of the geocoding operation, while creation of the person universe was deferred until the AREX address universe was trimmed and corrected via the several AREX specific operations designed to produce as accurate a list of addresses as possible. Before the post-processing phase of AREX (where the final population tallies were

produced), an AREX person universe was created by extracting appropriate StARS database records in the following manner:

- All persons ***ever*** associated with an AREX address were included in the AREX person universe file. Such a universe enabled the capture of "movers" into, out of, and within the AREX test site counties.
- By matching the address identifiers present on the AREX Address File against the AREX Pointer File, the unique record identifiers (UIDs) from original source input files were identified for extraction from the StARS database. This UID file was indexed and matched against the Linked Person Files (which contained addresses) in the StARS database.
- UIDs matched on the indexed file and the Linked Person File identified the SSN records with at least one address in the AREX test sites. An SSN list file was created and re-matched against the Linked Person File to select all records for any SSN on the list. In this fashion, every record (to ensure every address) for a given SSN was included in the AREX Person Universe File.
- The inclusion of all records (even non-AREX address records) for an SSN in the AREX Person Universe File ensured that only a person's best address was selected. Thus, if the best address for any person record from among the AREX person universe file was determined not to be within the AREX test site, the person record was flagged "out of scope" to ensure the person was not counted in the population tallies for the AREX test site.

4.1.10 Successes and Shortfalls in developing the StARS/AREX database

Successes achieved in developing the StARS/AREX database include:

- **Magnitude of Task**

  Creation of the StARS database was a prerequisite to the conduct of the Administrative Records Experiment. Regardless of the AREX population tally method employed, identification of the AREX universe was dependent upon an address list presumed to be within the AREX test sites. The AREX Address File (AAF), and each subsequent iteration, served as the address list. The difficult task of re-linking the final address file with the correct person record with intermediate pointer files and unique record identifiers was successfully accomplished.

  The myriad intermediate processes (and operational files) required during address unduplication, SSN verification, person processing, CPR building, and ultimately AREX person universe identification, encountered technical problems magnified by the sheer volume of records that required processing. More than 795 million address records from the initial source files were reduced to approximately 136 million unique address records on the StARS Master Housing File. AREX test site address records (approximately 1.3 million) were extracted from StARS. More than 875 million person records were matched against the Census Numident (which contained more than 396 million records) and the Person Characteristics File (also 396 million records), which resulted in approximately 279 million verified person records residing on the CPR. The SSN verification and search process also yielded approximately 30 million unverified person records. The AREX person universe

consisted of approximately 2.8 million records derived from StARS CPR processing. All told, more than 2.4 ***billion*** records were processed before arriving at the final AREX population tallies. The success of the AREX process demonstrated the capability of the Census Bureau to process a huge volume of records through a series of complex data processing steps with limited resources.

- **Multiple Addresses and Unduplication**

Two issues surfaced here. The first was the problem of Selective Service System input records that allowed for reporting two addresses on a single input record (current or permanent address). Both addresses were carried forward throughout the StARS processes until the address and person records were re-linked prior to creation of the CPR. The multiple addresses made identification of a best address difficult and created record count problems during each phase of address processing. During creation of the CPR, ARR staff determined[8] that use of the current address would most closely simulate a census operation — thus, the problem of multiple addresses for a single person record was eliminated.

The second issue involved the unduplication of records throughout the various processes. During the initial unduplication, prior to forwarding the records for geocoding, ARR staff erroneously unduplicated records based on an exact match of the full 9-digit ZIP Code. The inconsistent use of a ZIP + 4 code on administrative records was not fully considered in the application of the unduplication rules. As an example:

127 Oak Street, Dayton, Mo, 63901-1234, and

127 Oak Street, Dayton, Mo, 63901-0000

were considered two unique addresses. The error was overcome during construction of the Master Housing File, as long as the addresses in question were geocoded with a TIGERLINE ID. Under normal circumstances, both records in question would be assigned the identical TIGERLINE ID, and would therefore, be assigned an identical Housing Unit Identifier (HUID) to replace the actual street address. The records were then unduplicated based on exact match of the HUIDs. If the addresses were not geocoded, however, both records would continue to be treated as unique addresses on the MHF. Such occurrences were unlikely for city-style addresses as both Code-1 and the geocoding operations made "equivalent" matches on ZIP Code, street name, and combinations thereof. Where the problem surfaced for non-geocoded and noncity-style addresses, the issue was deemed less critical since AREX required geocoded addresses for inclusion in the tallies.

The following shortfalls were observed in developing the StARS/AREX database:

- **Name Entry Problems with the IRS W-2/1099 File**

The overwhelming majority of unverified SSNs came from the IRS W-2 / 1099 file where name entry data was often encumbered with commercial firm names such as banks, accountants, tax attorneys, etc. Such data was not unexpected given the nature of the tax data reporting inherent in the IRS W-2 / 1099 form itself. The problem was compounded by the fact that the social security numbers for spouses were often switched or a parent's name is

[8] Use of the Selective Service "current" address was based on the fact that more than 80% of the records reflected identical current and permanent addresses. SSS addresses contributed less than 1% to the overall number of addresses in the StARS database, thus marginalizing the effect of using only the current address from the file.

reported (present) on a child's SSN record. Many of the words in the name field were not recognized by the name standardizer during the person edit phase of StARS processing.

More than 25 million of the approximately 30 million records that did not verify during the SSN Search and Verification process were IRS W-2 / 1099 records. As with the address standardizer, the Census Bureau's name standardizer is dynamic and subject to operator input controls to achieve desired results. A consistent and methodological use of the name standardizer should achieve better results.

- **Dependent Names on the IRS 1040 File**

The IRS 1040 file, while reporting additional SSNs as dependents on the input file, provided only the first four characters of the identified dependent's last name. Lacking other demographic data (gender, date of birth, first name, etc), verification of the IRS 1040 dependents was an extremely difficult task (see remarks above on use of the IRS 1099 file). In fact, of the approximate 30 million of unverified SSNs, over ten percent (~ 3.1 million) were IRS 1040 dependent records.

- **Census as the "Gold Standard Assumption"**

The act of creating an "extract" for AREX purposes required implementation of demographic and address selection rules somewhat at odds with the StARS selection rules. The primary difference between AREX and StARS regarding address selection was an expressed requirement to find an address identifiable on a "piece of ground" within the AREX test site. To accomplish this goal, the StARS '99 and AREX address file processing address selection logic always deferred to a geocoded address over a non-geocoded address. This preference was driven by the fact that AREX required tabulation block geography in order to be included in the population tallies (regardless of method). As with the Decennial Census, the Decennial Master Address File (DMAF) served as the "gold standard" master address list for the AREX operation to satisfy the requirement to assign a block code for addresses in order to include persons in the block-level tallies. The issue of duplicate, multiple, and surviving MAFIDs on the DMAF created some data processing difficulties in ultimately selecting the "best" address with reasonable assurance. The issue may have contributed to more than a few erroneous "best" address selections to ultimately appear on the final AREX Address File. By default, population tallies where persons would be "assigned" to such addresses (block tallies) may be in error to a minimal degree. Similarly, deference to selection of a geocoded address for a given person record may have overlooked a more current address that was not geocoded. Post office box addresses, rural route addresses, and property name addresses may all be "better" addresses in many situations for certain person records. Further research and analysis into this problem are required to fully assess the impact on the viability of the final population tallies for both the Top-down and Bottom-up methods. See the AREX Household Evaluation for more information regarding match rates and housing unit totals.

4.2 Operational Components of AREX

Operational components of AREX were conducted on records contained within the five test site counties. These operations consisted of:

1. Master Address File Geocoding Office Resolution (MAFGOR),
2. Computer matching the AREX address to the Decennial Master Address File (DMAF),
3. Clerical review of the results of matching the AREX Address File to the DMAF,
4. Field Address Verification,
5. Request for Physical Address, and
6. Tabulating the results.

4.2.1 MAFGOR

As part of the creation of StARS, addresses were unduplicated across source files and split into three-digit ZIP Code files. The files were sent to Geography Division (GEO) for computer geocoding with the Maryland and Colorado files (that included the test sites) given priority.

During the computer geocoding, GEO selected and flagged addresses in the AREX 2000 test sites. Two different approaches were taken depending on whether the address was geocoded. If geocoded, an address was flagged as being within the test sites if the county/block codes fell within the test sites. If the address was not geocoded, the address was flagged as possibly being in the test site based on ZIP Code.

After the selection of the test site records, addresses that were not computer geocoded, but were within a test site ZIP Code, were subjected to clerical resolution through MAFGOR. Addresses eligible for MAFGOR were formatted and sent to the Regional Census Centers in Denver and Philadelphia where clerical geocoding of the addresses were attempted and the results keyed. After the MAFGOR results were keyed, the records were returned to GEO where the results of the MAFGOR were updated to the geocoded file. The file was then sent to PRED to update the AREX Address files prior to the computer matching of the file to the DMAF.

4.2.2 Computer Matching of AREX Records to the DMAF

- **Description of the Computer Match Process**

  The objective of the computer match operation of AREX was to determine the extent and nature of matches between addresses from administrative records source files and eligible addresses from the Census Bureau's Decennial Master Address File (DMAF) in support of the Bottom-up method. The concept of the Bottom-up method is to start with a known list of residential addresses (in this case the DMAF), match the administrative records to such a list and reconcile any non-match cases.

  The AREX file used for this process was the iteration containing geocode information from the computer geocoding and MAFGOR process. Prior to the matching process, address fields of the AREX file were standardized using the Geography Division's address standardizer software program. The file to which the AREX file was matched consisted of a

list of addresses on the MAF whose current county code showed the address to be within one of the five AREX test site counties.

The matching process consisted of running AutoMatch, a commercial software package to match the addresses from the two files. AutoMatch was run in three passes to match both geocoded and ungeocoded city-style addresses. Matching results were based on parameter settings established by PRED analysts. The final results were divided into matches; possible matches; non-matches and matches to duplicate DMAF addresses.

Not all addresses on the administrative records file were sent to the AutoMatch process. To most accurately match the addresses, the match was limited to addresses with a standardized street name, a standardized property description or both. Excluded from the matching process were non-*standardized* addresses, standardized post office or box addresses, standardized post offices, rural route addresses, and undefined addresses.

- **Results of the Matching**

Table 6 shows the results of the computer matching by number of addresses forwarded to the computer match.

Table 6. Computer Match Results

Test Site County	AREX Records to Computer Match	DMAF Records[1]	Addresses Matched	% of Addresses Matched[2]	Possible Matched Records	Non-Matched Records	Duplicate Matches
Baltimore City Maryland	303,003	329,797	234,360	77%	870	67,646	127
Baltimore County Maryland	353,278	323,074	290,875	82%	1,264	60,847	292
Douglas County Colorado	65,294	65,027	52,574	81%	1,700	10,926	94
El Paso County Colorado	226,110	208,416	178,279	79%	2,900	44,683	248
Jefferson County Colorado	241,987	259,366	201,288	83%	7,501	32,982	214

1. DMAF records include all address types
2. Some administrative records matched to more than one address in the DMAF, each of which might have had subtle differences. When this occurred, addresses were flagged as having duplicate matches. The duplicates were resolved later in the AREX operation where the best address was selected based on pre-selected criteria.

- **Lessons Learned from the Computer Matching Process**

During the matching process, there was an ongoing analysis of results and subsequent AutoMatch parameter adjustments to ensure optimum match rates of the addresses. In spite of the extensive analysis, however, it was found that the accuracy of the computer match can be improved with a clerical review or follow-up field operation specifically looking a the possible matches and non-matched cases. The subsequent AREX clerical review process supported the notion that it was important to follow up a computer match of administrative records with some type of clerical review or other type of match reconciliation process. As

shown in the following section on clerical review, many addresses determined to be possible matches in the computer match were ultimately matched during the clerical review process.

A factor to consider when matching administrative record addresses is the vintage of address information in the file to which it is matched. The AREX address records contained data that were from 1999 and earlier. Analysis of matching results should consider possible address changes made between the vintage of the administrative record address and the vintage of data of the DMAF to which the file is matched.

A more consistent method of address standardization should improve the overall match rate. Throughout the course of creating the StARS database and subsequent iterations of the AREX address file, the Geography Division's address standardizer was employed. The dynamic nature of the standardizer software program and the flexibility of operator control during its application most likely contributed to inconsistencies and variances that led to erroneous matches (and non-matches as well). A dedicated, fixed version of the standardizer should be used throughout the entire administrative records census process. Although difficult to quantify, the application of a fixed version of the standardizer along with prescribed operator control methodologies should improve the overall match rate during the computer matching operations. Improving the computer match rate would reduce the number of address records requiring clerical review.

The AREX address files were matched against a version of the DMAF extract file. Multiple MAFIDs assigned to a single address and duplicate MAFIDs assigned to multiple addresses contributed to the difficulty in classifying an address as matched, non-matched, or possibly matched. During the later re-match to the DMAF the multiple and duplicate MAFID issue compounded matching effort inconsistencies - probably due to the Census Bureau's methodology and audit trail for identification and retention of "surviving MAFIDs" on the DMAF. We recommend further research into the impact of DMAF rationale for retaining duplicate and multiple MAFIDs on the file.

4.2.3 Clerical Review of the Matching Results

- **Description of the Clerical Review Process**

Following the computer match, a clerical review was conducted by the staff at the National Processing Center. The clerical review process supported the AREX Bottom-up method in that administrative records are assigned to individual housing units and inconsistencies between the addresses must be resolved. The main purposes of the clerical review were to:

1. Review all addresses designated as possible matches by the computer match and determine whether these addresses should be coded as a match to each other, a match to other addresses or a non-match. The search for a matching address was first done based on addresses in the same ZIP Code. A DMAF listing, sorted alphabetically by street name was also provided for researching clusters of unmatched administrative record addresses that could not be found in the ZIP Code of the DMAF listing provided.
2. Review all AREX addresses coded as non-matches by the computer match and determine if a clerical match can be made to the DMAF.

3. Review the non-matching AREX addresses and determine if the addresses are incomplete, contain extraneous data or have any other unusual characteristics that made the address unsuitable for field address verification.

- **Results of the Clerical Review**

The clerical review process made final match determination for all possible and non-matched addresses and identified cases that the computer could not match but a clerical reviewer could. AREX non-matched addresses were also reviewed to flag addresses not eligible for field address verification because they were incomplete or contained inappropriate information that could not be verified (APO addresses, foreign addresses, in care of, etc).

Due to program deadlines regarding the Field Address Verification operation, there was a point during the keying of clerical review results where a cutoff was made and materials were produced for the FAV sample of addresses (discussed in more detail in section 4.2.4). Because of this cutoff, not all of the match status results were keyed. Keying of the remainder of the clerical review results was accomplished later in the AREX program and records were flagged to distinguish the first keying from the second keying.

The second keying added an additional 11,397 matched records to the database from the point at which the FAV sample was defined. Additionally, 77 addresses defined as matches after the first keying, were found to be non-matches in the later stages of the clerical review with the records updated accordingly in the second keying. The results of the clerical review after the second keying are shown in Table 7.

Table 7. Clerical Review Match Results

Test Site	Number Records to Computer Match	Number Matched by Computer	Number Matched after Clerical Review	% of Records matched by Clerical Review
Baltimore City Maryland	303,003	234,360	241,557	2%
Baltimore County Maryland	353,278	290,875	302,332	3%
Douglas County Colorado	65,294	52,574	56,592	6%
El Paso County Colorado	224,105	178,279	188,866	5%
Jefferson County Colorado	241,987	201,288	214,298	5%

- **Lessons Learned from the Clerical Review Process**

The original AREX plan called for PRED staff to do the clerical review of the unmatched and possible-matched records. Based on assuming an increased workload for the AREX Field Address Verification, PRED contacted DSCMO, requesting NPC staff to conduct the clerical review. PRED trained approximately 25 reviewers to evaluate the possible matches of AREX addresses against the DMAF and make a match/non match determination for the address. The reviewers attempted to match the non-matched AREX addresses to the DMAF. Printouts of the addresses were sent to NPC for clerical review. After the clerical review, the sheets were returned to PRED for QA and keying. Although all addresses were eventually reviewed and the results keyed, geographically separating the components of the operation created additional coordination and deadline challenges.

In future clerical review operations, administration of the process would be eased by having all components of the operation conducted in a central location. Although having NPC do the clerical review assisted PRED by freeing up PRED staff for other functions of AREX, the geographical separation of components created an additional dimension of the operation.

More detailed practice examples were needed in the training. The clerks needed to develop a clearer understanding of how matches and non-matches were defined in AREX.

A better understanding of how the DMAF was developed might have improved our implementation of clerical review**.**

To maximize the effectiveness of a clerical review, we recommend that reviewers be thoroughly trained in the unique intricacies of addresses derived from administrative records. This background could assist the reviewers to better determine the match/non-match status of the address. It should be noted that only minor problems were created by the two waves of keying results from the clerical review. However, in an effort to produce as clean a list as possible of addresses eligible for a FAV operation, greater control should be exercised over any clerical review operation.

A post-production clerical review should also be considered for future administrative records experiments as an evaluation tool or audit-trail validation system. Again, the application of precise definitions for what constitutes a match, non-match, or possible match are the critical elements of any clerical review operation**.**

4.2.4 Field Address Verification (FAV)

- **Description and Purpose of the FAV**

The Field Address Verification Operation was implemented to check the validity of addresses that remained unmatched to the DMAF following the computer matching and clerical review. To minimize the amount of field work required, the assumption was made that any non-matched DMAF addresses were in fact, valid and existent because of the numerous operations that went into the building of the DMAF. As a result, only non-matched administrative record addresses were eligible for the FAV. The purposes of the FAV included:

- Verifying the physical existence or nonexistence of non-matched AREX 2000 Test Site addresses.
- Correcting erroneous address field values.
- Identifying addresses meeting unique conditions - such as a duplicate of another address.

The operations conducted in the FAV were quite different from the planned FAV operation. The original plan called for the Technologies Management Office (TMO) to create an input file for production of address listing pages for all non-matched AREX addresses and to then ship the pages to the appropriate Local Census Office (LCO), which was to conduct the address verification operation. The LCOs and Regional Census Centers (RCCs) were to send the listings to the NPC for keying upon completion of the operation. The plan was modified due to the decennial commitments of Divisions whose participation was needed to produce the listing pages and conduct the FAV. PRED redesigned the operation to enable use of its

own resources to conduct the FAV. The primary impact was that only a sample of addresses were selected across the test sites for field verification.

- **Results of the FAV**
  - ***Defining the Sample***

    After the computer phase of address matching, the universe of addresses eligible for Field Address Verification was first restricted to geocoded, city-style addresses within the AREX 2000 test site counties. The universe was further restricted to exclude some AREX 2000 test site ZIP codes that belonged to three colleges, a medical center, and an Air Force base in the belief that few or no residential addresses existed in them.

    With the redesign of the FAV operation, the number of addresses to be verified was based on a stratified cluster sample of unmatched, city style addresses. The sample resulted in 6,644 addresses being flagged as part of the FAV operation. Table 8 displays the number of records eligible and selected for each test site state.

Table 8. Selection of FAV Addresses

Test Site State	FAV Eligible Addresses	Addresses Selected for FAV Sample
Colorado	57,333	3,730
Maryland	96,202	2,914
Total	153,535	6,644

For each address selected as part of the sample, an address listing page was printed. A sample listing page is displayed in Figure 3. Each listing page contained the address information and a series of yes/no questions regarding the address. The lister was responsible for answering each of the yes/no questions and provide amplifying remarks or comments when required.

Is this address…	Yes (1)	No (2)
1)…found in the search area?	1)	1)
2)…residential?	2)	2)
3)…commercial?	3)	3)
4)…a special place or group quarters?	4)	4)
5)…geocoded to the correct block?	5)	5)
6)…found in the county shown above?	6)	6)
7)…correct as shown above?	7)	7)
8)…shown as a multi-unit but is actually a single unit?	8)	8)
9)…shown as a single unit but is actually a multi-unit?	9)	9)
10)…a duplicate of another AREX address?	10)	10)
11)…unresolvable due to insufficient information?	11)	11)

Figure 3. Depiction of FAV Listing Page Questions

- ***The Listing Operations***

  PRED solicited volunteers from within and outside the Division to conduct the FAV. The twenty volunteers were divided into two teams - one team for Colorado addresses and the other team for Maryland addresses. To prepare the listers for the field operation, a two-day training seminar was conducted. In addition to the classroom training, teams were given a listing assignment for a residential area close to the Census Bureau. Results of the field training were reviewed and debriefed prior to certifying the volunteer listers. Listing pages were grouped by block and assigned to the team leader for the appropriate area and a set of maps were produced for each area. Team leaders tasked each team member with a set of blocks to list. Two weeks were allotted to complete the field work. All team members were issued cell phones with which to communicate with each other and to communicate with individuals at PRED designated to support field operations.

  - Listers were tasked to assess the following from each of the listed addresses:
  - Does the address exist?
  - Does the address require a correction?
  - Is the address a duplicate?
  - Is the address a special situation (for example, commercial/nonresidential, special place/group quarters, etc)?

- ***Processing the Field Listing Results***

  After the field work was completed, listing pages were reviewed for completeness and accuracy before sending them to NPC to be keyed. NPC keyed the information from the listing pages and returned the listing pages and keyed files to PRED. PRED staff then reviewed each of the listing pages and annotated a 5-digit status code on the page. The code categorized the type of activity about the address that was shown on the listing page. Status codes were also used to categorize if the address was valid. In some instances, the address was valid as listed (without changes). There were instances where corrections could be made to the address to make the address valid; yet other cases where even with changes, the address was not valid. Attachment 3 to this document defines the status coding and valid/invalid address criteria. These codes were keyed into a separate file and later added to the AREX Address File. Table 9 shows the results of the listing in terms of the valid status of the addresses.

Table 9. FAV Results

Test Site County	Number of Addresses Sampled	Percentage Valid as Listed	Percentage Valid After Lister Corrections
Baltimore City Maryland	1513	15%	24%
Baltimore County Maryland	1401	10%	26%
Total Maryland	**2914**		
Douglas County Colorado	1226	6%	38%
El Paso County Colorado	1344	10%	22%
Jefferson County Colorado	1160	6%	44%
Total Colorado	**3730**		

Note: The percentages listed above are based on their proportionality in the FAV sample and are not weighted relative to their probability of selection from the FAV eligible address universe. For a detailed discussion of the FAV Sample Selection process, see the Consolidated AREX Evaluation Report.

Of the addresses in the FAV sample, 84 were found to be commercial mailbox addresses.

Of particular interest in Table 9, are the results shown in column three depicting the percentage of addresses found to be valid as listed. Although the addresses were valid as listed, the fact they were included as part of the FAV sample because they did not match to any address in the DMAF either in the computer match or in the clerical review operations. The non-match status could have been due to anomalies of the matching process or the currency of data within either the AREX or DMAF files.

- ***Applying the Results of the FAV***

  Although only a sample of addresses were sent for field verification, in order to complete AREX address file processing, an assessment had to be made on the status of FAV eligible addresses that were not part of the sample. To do this, a FAV estimation model was developed. The FAV estimation model is a logistic regression model defined by the following formula:

$$\frac{1}{(1+e^{-(\beta_0+\beta_1 Var_1+\beta_2 Var_2+\ldots\ldots\beta_n Var_n)}}$$

  β = estimated model coefficient – static after model has been determined

  $Var_1, Var_2 \ldots\ldots Var_n$ = independent variable specific to each address being modeled (addresses that are FAV eligible, but not in the FAV sample)

  All FAV eligible addresses that were not selected as part of the sample were designated as either valid or invalid address based on the application of the FAV logistic regression model. The purpose of the FAV logistic regression model is to predict the validity of a FAV eligible address. An address is considered valid if it is a

non-group residential structure not already shown on the Master Address File. To make these predictions, we built a logistic regression model that incorporated several address characteristics. We calibrated this model using the results of Field Address Verification. In order to estimate the reliability of the model, we took half samples of FAV-included addresses, estimated model parameters from this half sample, and used the estimated parameters to predict the known valid-status of the other half of the sample. Based on 200 replications with randomized half-samples, we estimate that the model accurately predicts address validity 69 percent of the time. As this result is based on half-samples, we expect actual results to be somewhat better than this as they will be based on a sample twice the size.

- **Lessons Learned from the FAV**
  - ***Inexperienced Listers doing the Field Work***

    Because PRED assumed responsibility for the FAV, volunteers with little to no field experience were doing the field listing operation. To minimize the impact of this, PRED decided not to use listers in the traditional role of assigning action codes but rather to collect information about the address for later analysis and assignment of the action code. In the revised FAV, listers answered 11 questions about the property from which the action code (called status code in this operation) was later assigned. This modification worked well in minimizing the mistakes made by inexperienced listers and created a collateral benefit of collecting detailed information about the address for further research and analysis.

  - ***Sample versus Full FAV***

    A key impact of the revised FAV operation was that only a sample of addresses were verified in the field. Although great pains were taken to develop the sample and estimation formulas, there is no way of assessing how well the samples actually represented the FAV eligible universe. Further research should be dedicated to study this issue.

    To ensure the best possible correlation to decennial census operations, future administrative record experiments should do a field listing on all eligible addresses using the same skilled listing personnel that are used for decennial operations.

  - ***Commercial Addresses***

    During creation of the StARS database, address records were matched against a software product to identify potential commercial addresses. The product used was the American Business Information (ABI), Inc. database file of commercial addresses (more than ten million) based on national telephone directories (both yellow and white pages). The identification and removal of commercial addresses from the AREX address files is critical to create an accurate population tally. Budgetary restraints precluded purchase of the ABI residential file. The use of both files (commercial and residential) would have improved the accuracy of commercial address identification and (perhaps more importantly) improved the accuracy of the FAV eligible address list as well. We recommend additional software products be evaluated for suitability in the conduct of an administrative records census.

4.2.5 Request for Physical Address Operation

The AREX 2000 Request for Physical Address Operation collected physical addresses (house number and street name) for individuals with a Post Office Box or other noncity-style address from the administrative records source files. Major components of the operation were to:

1. Create an address file from administrative records where the mailing address was a Post Office Box or noncity-style address.
2. Design a form and mail it to the addresses, requesting a physical address.
3. Clerically geocode the physical addresses on the forms to state, county and block.
4. Key addresses and geocode information to a file for further analysis.

Based on low response rates, it was decided to not incorporate the results of the Request for Physical Address operation into the AREX operation but to create a separate analysis of the operation. Details of this operation are included in the AREX 2000 Request for Physical Address Evaluation.

4.2.6 Rematching to the DMAF and Producing the Baseline Tallies

A final match of the AREX addresses was made to the DMAF for the purpose of transforming the collection geography to tabulation geography. Because the AREX addresses were initially geocoded to collection geography, it was necessary to translate the collection geographic codes into the tabulation geographic codes so that the comparisons to Census 2000 tabulations could be made. The taking of the census spans approximately a two-year period, including the address list building phase. The geographic framework going into the census is called collection geography. Prior to tabulation of the final counts, changes must be incorporated to reflect boundaries in effect on January 1, 1999. This final geographic framework is called "tabulation" geography.

During the rematch, a problem with duplicate and multiple MAFIDs present on the DMAF resurfaced (see recommendation section). The dynamic nature of the DMAF requires that MAFID assignments be continually updated from decennial census operations. Thus, the number of duplicate and/or multiple MAFIDs for a given address may have changed since the first computer match. The impact on correctness of tabulation block assignments just prior to generation of both the Top-down and Bottom-up tallies is difficult to assess. Further research into the multiple/duplicate MAFID issue relative to the DMAF should be undertaken.

The final step in the AREX operation, prior to post processing was to create tallies for both the Bottom-up and Top-down methods of the experiment. The purpose of the tallies was to serve as a basis of comparison of an administrative records census to a conventional census tally.

The Top-down tallies are drawn from AREX records at the census block level or above. The tallies for the Top-down method are shown in the following table.

Table 10. Top-down Method Population Tallies

Test Site County	AREX Population	Census Population	% of Census Population
Baltimore City Maryland	570,648	651,154	88%
Baltimore County Maryland	696,183	754,292	92%
Douglas County Colorado	148,270	175,766	84%
El Paso County Colorado	456,891	516,929	88%
Jefferson County Colorado	473,495	527,056	90%

Note: Top-down tallies include Group Quarters (GQ) addresses if, during the rematch to the DMAF, the administrative record was matched to a DMAF GQ address.

The Bottom-up tallies are drawn from AREX records at the household level and above. The Bottom-up tallies are shown in Table 11 below.

Table 11. Bottom-up Method Population Tallies

Test Site County	AREX Population	Census Population	% of Census Population
Baltimore City Maryland	661,561	651,154	102%
Baltimore County Maryland	745,893	754,292	99%
Douglas County Colorado	170,102	175,766	97%
El Paso County Colorado	509,597	516,929	99%
Jefferson County Colorado	508,254	527,056	96%

Note: Bottom-up tallies contain GQ addresses that were: (1) identified in the administrative record only, (2) identified in the DMAF address to which the AREX address was matched, or (3) identified as a Census only address.

4.2.7 Successes and Shortfalls of Producing the Tallies

The process to create the tallies worked well. Any issues regarding the quality of data that was used to generate the tallies, are covered elsewhere in this report.

5. RECOMMENDATIONS

This section of the report discusses lessons learned from the operation and provides recommendations regarding initiatives that may apply to Census Bureau use of administrative records in future census related activities.

5.1 Improve the computer matching and rematching processes

Computer address matching parameters must be further evaluated for accuracy and relevancy to the address matching task at hand, as many addresses classified as possible matches by the computer were deemed to be matched during the clerical review process. A second match to the DMAF was conducted for the primary purpose of assigning tabulation block geography to the AREX address files to facilitate the reporting of population tallies at the block level (as in the decennial census). The impact on correctness of tabulation block assignments just prior to generation of both the Top-down and Bottom-up tallies is difficult to assess. An evaluation should be conducted to determine the effectiveness of the rematch to the DMAF process. The dynamic nature of the DMAF requires that it be continually updated from decennial census updates. Thus, duplicate and multiple MAFIDs for a given address may have changed since the first computer match. During the rematch, a problem with duplicate and multiple MAFIDs present on the DMAF resurfaced (see also recommendation 5.2).

5.2 Evaluate the impact of multiple MAFIDs on the DMAF

We recommend further research on the impact of retaining duplicate and multiple MAFIDs on the DMAF file should be pursued, particularly as related to the match to administrative record files. The AREX address files were matched against a version of the DMAF extract file to establish matched and non-matched addresses for the Bottom-up method. Multiple MAFIDs assigned to a single address and duplicate MAFIDs assigned to multiple addresses contributed to the difficulty in classifying an address as matched, non-matched, or possibly matched. During the later re-match to the DMAF to transform "collection' geographic codes to "tabulation" geographic codes the multiple and duplicate MAFID issue compounded matching effort inconsistencies - possibly due to the Census Bureau's methodology and audit trail for identification and retention of "surviving MAFIDs" on the DMAF.

5.3 Improve the availability of source data for the under 18 population

Administrative Records Research should continue to pursue coverage improvements via additional file acquisition. Expanding coverage of existing files should also be pursued in an attempt to improve coverage of certain segments of the population — particularly dependents on the Internal Revenue Service files and the under age 18 population segment nationally. Improving race information on administrative record files should also be pursued.

Administrative record source data for the under age 18 population presents a particular problem for conducting an administrative records census. Although the IRS 1040 file can list up to four dependent children for each tax return, the lack of demographic data in this file is an inhibitor to successful verification of this segment of the population. The IRS W-2 / 1099 file (which could enhance the IRS 1040 data by providing a full name with an SSN) requires numerous file-specific edits and "work-arounds" in an attempt to extract "clean data" for verification and matching purposes. Oftentimes, the IRS W-2 / 1099 file simply "clutters the landscape" with

regard to best data selection for output to the Composite Person Record. Again, the lack of demographic data on the IRS W-2 / 1099 file compounds the verification problem.

We recommend the search for another file source to better capture the under 18 population. Public school enrollment files or school lunch data files may be the best source of such data. However, because these programs are generally controlled at the state or local government level, compatibility and standardization of files from the various states (even if obtainable) could be a major deterrent to efficient processing of the data. In addition to more source files, increasing the value of information available from existing source files is also a possible approach. Options may include seeking access from IRS records for all dependents and modeling additional children using total number of exemption data available on the tax return.

5.4 Evaluate the effectiveness of computer models used in the experiment

We recommend an evaluation of the effectiveness of the models used for the FAV. Because the AREX FAV was reduced in scope to a sample, two models were established to compensate for the less than full field evaluation. One model was used to select the best sample of addresses to be sent to FAV, the other model was used to apply the results of the FAV to an estimation relative to the non-sampled addresses. While independent quality assurance was employed throughout the development and use of the models, it is important to do further analysis of the effectiveness of the models, particularly since the final tallies and results of the experiment are influenced by the models.

5.5 Conducting further research on address selection

We recommend an evaluation be made on the effectiveness of the process used to select the best address during StARS/AREX processing. Address selection is a linchpin process in the conversion of administrative record source data to a format acceptable to generate census tallies. As such, a system that maximizes available information to select the best address is critical. One method of evaluating this factor might be to match the selected address to the census to validate the effectiveness of the rules used in the process. A more thorough assessment of the StARS and AREX address selection rules used to determine a person's "best address" should be pursued.

5.6 Conduct a full-scale field address verification

We recommend the next administrative records experiment complete a full-scale field address verification operation. In AREX 2000, there was only enough time and resources to field verify a sample of the addresses that did not match to the DMAF. In addition, we assumed that DMAF non-matches were "truth" because of census operations to build and confirm the DMAF. A more thorough approach would be to field check both the administrative record and census non-matches. We used the results of the sample to build a model for predicting how many of the unverified non-matches were actually valid addresses. Our experience suggests the ability to predict the number of valid addresses from a model is extremely limited. We believe more precise results can be obtained from larger field address verification.

REFERENCES

Aguirre International (1995). *Public Concerns About the Use of Administrative Records*. Unpublished document available from the U.S. Census Bureau, July 12, 1995.

Alvey, Wendy and Scheuren, Fritz (1982). Background for an Administrative Record Census. *Proceedings of the Social Statistics Section*, Washington DC: American Statistical Association, 1982.

Buser, Pascal, Huang, Elizabeth, Kim, Jay K., and Marquis, Kent (1998). *1996 Community Census Administrative Records File Evaluation*. Administrative Records Memorandum Series # 17. Washington, DC: U.S. Bureau of the Census.

Bye, Barry (1997). *Administrative Record Census for 2010 Design Proposal*. Washington, DC: United States Department of Commerce.

Wiley Conklin, Joseph (2001). *Administrative Records Experiment in 2000 (AREX 2000), Evaluation 2: Processes, Initial Draft, December 31 2001*.

Czajka, John L., Moreno, Lorenzo, Schirm, Allen L. (1997). *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population*. Washington, DC: Internal Revenue Service.

Edmonston, Barry and Schultze, Charles (1995) *Modernizing the U.S. Census*, National Academy Press, Washington DC.

Gellman, Robert (1997). *Report on the Census Bureau Privacy Panel Discussion*. Unpublished document available from the U.S. Census Bureau, June 20, 1997.

Knott, Joseph J. (1991). *Administrative Records*. Memorandum for Distribution List, Bureau of the Census, Washington DC, U.S. Bureau of the Census, November 12, 1991.

Myrskyla, Pekka (1991). Census by questionnaire—Census by registers and administrative records: The experience of Finland. *Journal of Official Statistics*, 7:457-474.

Myrskyla, Pekka, Taeuber, Cynthia, and Knott, Joseph (1996). *Uses of administrative records for statistical purposes: Finland and the United States*. Unpublished document available from the U.S. Census Bureau.

Pistiner, Arona, and Shaw, Kevin A. (2000). *Program Master Plan for the Census 2000 Administrative Records Experiment (AREX 2000)*. Administrative Records Research Memorandum Series #49. U.S. Census Bureau.

Singer, Eleanor, and Miller, Esther (1992). R*eactions to the use of Administrative Records: Results of Focus Group Discussions*. Census Bureau report, Center for Survey Methods Research, August 24, 1992.

Zanutto, E. (1996). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. Presentation to the U.S. Bureau of the Census, 8/26/96.

Zanutto, Elaine, and Zaslavsky, Alan M. (1996). *Modeling census mailback questionnaires, administrative records, and sampled nonresponse followup, to impute census nonrespondents*. In Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Zanutto, Elaine, and Zaslavsky, Alan M. (1997). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. In Proceedings of the U.S. Bureau of the Census Annual Research Conference. Washington, DC: U.S. Census Bureau.

Zanutto, Elaine, and Zaslavsky, Alan M. (2001). *Using administrative records to impute for nonresponse*. To appear in R. Groves, R.J.A. Little, and J.Eltinge (Eds), Survey Nonresponse. New York: John Wiley.

The following document list is a compilation of Planning, Research and Evaluation Division (PRED) technical specifications that define the process to create the Statistical Administrative Records System StARS) 1999 and the various operations of the Administrative Records Experiment (AREX) 2000. Read publication year (parenthesis), document title, and document catalog number (parenthesis).

StARS 1999

Address Editing

(1999), Medicare Address Editing Programming Specification (MCAR9901-00)

(1999), HUD TRACS Address Editing Programming Specification (TRAC9901-00)

(1999) Selective Service Address Editing Programming Specifications (SSSX9901-00)

(1999) Indian Health Services Address Editing Programming Specifications (IHSX9901-00)

(1999) IRS 1040 Address Editing Programming Specifications (10409901-00)

(1999) IRS 1099 Address Editing Programming Specifications (10999901-00)

Address Processing

(2000) StARS Address Processing Programming Specifications (StAR9902-00)

Personal Characteristics File (PCF) Development

(2000) The StARS PCF Models: Executive Summaries - Supplement

(2000) PCF Creation Programming Specifications (StAR9906-00*)*

Person Editing

(2000) StARS Person Edit Programming Specifications (StAR9904-00)

(2000) Split/Segment of Edited Person Files Spec Sheet (StAR9905-00)

(2000) Medicare Person Editing Spec Sheet (MEDB9903-01)

(2000) HUD TRACS Person Editing Spec Sheet (TRAC9902-00)

(2000) Selective Service Person Editing Spec Sheet (SSSX9902-00)

(2000) Indian Health Services Person Editing Spec Sheet (IHSX9902-00)

(2000) IRS 1040 Person Editing Spec Sheet (10409903-00)

(2000) IRS 1099 Person Editing Spec Sheet (10999902-00)

Person Processing

(2000) Database Development 2000 (STAR9907-00)

(2001) Creation of the Final CAF (STAR9908-01)

SSN Verification and Search

(2000) SSN Verification and Search Programming Specifications (StAR9903-02)

AREX 2000 Specifications

AREX Address File (AAF)

(2001) AAF File Master Layout (uncataloged)

(2001) Summary of Updates from AAF1 to AAF9 (uncataloged)

(2001) AREX Address File (AAF) Naming Convention (uncataloged)

GEOCODED File

(2000) Selecting and Flagging Test Site Records (ARXG0001-02)

(2000) AAF1 Specification (ARXG0002-00)

(2000) AAF2 and Processing DocuPrint Control Specification (ARXG0003-00)

(2000) AAF3 and Results of the Clerical Geocoding Specification (ARXG0004-00)

(2000) Creation of AREX Person Universe File (ARXG0005-00)

Request for Physical Address (RFPA)

(2000) Specification for Printing and Mailing RFPA Letter (ARXR0001-00)

(2000) Specification for Creating the Address File for Input to DocuPrint (ARXR0002-00)

(2000) Specification for Check-in and Check-out of the RFPA Letter (ARXR0003-00)

(2000) Specification for Keying Data from the RFPA Letters (ARXR0004-01)

(2000) Receiving the RFPA Keyed Letter Data Files (ARXR0005-01)

(2000) AAF7 and Final RFPA processing Specification (ARXR0006-00)

Computer Address Matching

(2000) AAF3-DMAF Computer Matching Specification (ARXM0001-00)

(2000) AAF4 and Results of Computer Match and Clerical Review Spec (ARXM0003-00)

Clerical Review

(2000) Clerical Review Instructions (ARXM0004-00)

(2000) AAF5B and the Second CR Keying Specification (ARXM0006-01)

Field Address Verification (FAV)

(2000) AAF5 and FAV Sampling Specification (ARXM0005-00)

(2000) FAV Address Selection and Printing of Listing Pages (ARXA0001-02)

(2000) FAV Listing Page Check-In and Batching Specification (ARXA0002-00)

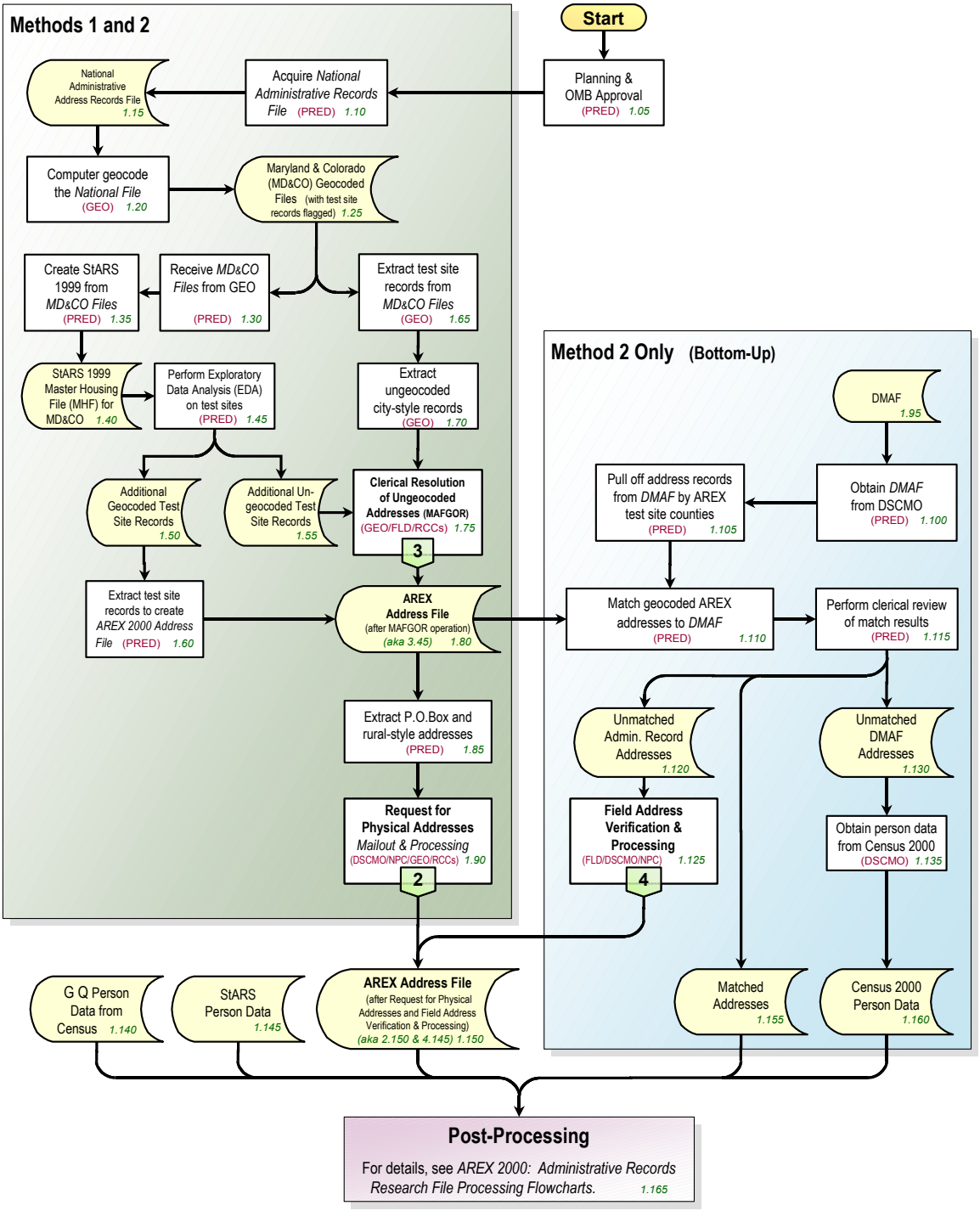
(2000) FAV Lister Instructions (ARXA0003-00)

(2000) Specification for Keying/Verification of the FAV Listing Pages (ARXA0004-01)

(2000) Specification for Receipt of the Keyed File from the FAV Listing Pages (ARXA0005-00)

(2000) Specification for Processing the FAV Keyed File and Creating AAF 6 (ARXA0006-04)

Attachment 1. AREX 2000 Implementation Flow Chart



Attachment 2. StARS Process Steps – Outline

The process steps outline that follows is a synthesized extract from pertinent StARS 1999 programming specifications. The outline is presented here to assist in understanding the complex nature (at a high level) of the operations required to build the StARS database. For a more detailed description of the processes, refer to the StARS specifications listed on the reference page (page 46) of this document. In outline format, the "dual-stream" processing steps in the creation of the StARS 1999 database are as follows:

1. **Edit and standardize address data** from the national-level source files.
   a. Combine all records and split resulting file into 1000 ZIP Code cuts in preparation for the Code-1 process.
   b. Pass records through Code-1 to standardize and "clean" the address data.
   c. Unduplicate the address records and create the GEO Extract File.
      1) Unduplicate on exact match of all address fields (full 9-digit ZIP Code).
      2) Extract file contains minimum number of data fields for TIGER coding.
2. **Edit and standardize person demographic data** from national-level files.
   a. Name edits and standardization designed to enable record matching, linking, and unduplication within the database once SSNs are verified.
   b. Split and sort records into Census Numident segments by Social Security Number (SSN) in preparation for SSN Search and Verification (S&V) phase of StARS.
3. **Verify and validate SSNs** by matching and comparing name data, date-of-birth data, and gender information against the Census Numident using AutoMatch.
   a. Pass unverified SSNs through "name/date-of birth search" phase using AutoMatch.
   b. Differing match cut-off scores and weights established for each source file.
   c. Use Census Numident data to fill missing demographic input data. Demographic data (other than name fields) for all IRS records derived from Census Numident.
   d. Person records now ready for re-link to the geocoded address records.
4. **Create** the **Master Housing File (MHF)** as follows:
   a. Pass the ABI commercial file through Code-1 and the address standardizer to format and "clean" commercial addresses.
   b. Unduplicate ABI file (exact match of parsed fields), and assign address type.
   c. Pass Geocoded files through the address standardizer to obtain parsed address fields in preparation for record unduplication.
      1) Assign address type based on standardized return fields.
      2) Unduplicate GEO files based on exact match of parsed fields within type.
   d. Merge unduplicated Geocoded file with unduplicated ABI file to identify and flag commercial addresses within each 3-digit ZIP Code file.

1) Assign a Housing Unit Identification Number (HUID).
2) HUID provides a numeric variable indicator to assist in selection of the best address for output to the final StARS database (the CPR).

e. Update the Master Pointer File (MPF) to enable address linkage back to original source files. MPF also reflects number of duplicate addresses associated with each address selected for retention on the MHF.

f. Merge the MHF and MPF and split resulting file back to original source cuts.
1) Select only the "current" address from Selective Service Records
2) Merge split files with source Proxy Files to append proxy addresses and create Enhanced Master Pointer File.

5. **Create Linked Person Files**

a. Use "direct access" method to link person records with Enhanced Master Pointer File.

b. UID variable identifies the correct EMPF source file to access for selecting required geographic data for inclusion on Linked Person File.

c. Link unverified SSN records in the same fashion.

6. **Create** the **Composite Person Record** (**CPR**) by selecting the "best record" from the Linked Person Files as follows:

a. Invoke address selection rules to **determine** the **best address** for the person records. Address selection rules follow:
1) Select the highest HUID category available.
2) Select a non-proxy address over an address with a proxy.
3) Select a non-commercial address over a commercial address.
4) Select the address based on source file priority as follows:
a) IRS 1040 record
b) Medicare record
c) Indian Health Service record
d) IRS 1099 record
e) Selective Service record
f) HUD TRACs record
5) Select most recent record based on the administrative record cycle dates.
6) Select first record read-in to the processing array for output to the CPR.

b. **Select** the **best race** based on the following rules:
1) If American Indian or Alaska Native is reflected on the IHS record, accept the value.
2) If an input value is blank or unknown – defer to the PCF.
3) Select the most frequent occurrence.
4) If tied among occurrences, defer to the PCF.
5) If record is from the "New SSN List," defer to the PCF.
6) If ties still occur, select first record read-in.

c. **Select** the **best** indicator of **Hispanic origin** based on the following rules:
   1) Most frequent non-blank observation (Numident value counted once).
   2) If ties occur, defer to the PCF.
   3) If the input value is blank, defer to the PCF.
   4) If record is from "New SSN List" and non-blank, output a positive Hispanic origin; if blank; output a blank value (SSN not on PCF).

d. **Select** the **best gender** based on the following rules:
   1) If a Selective Service record available, select "male" gender.
   2) Select most frequent occurrence, if no Selective Service record available.
   3) If ties occur among the observations, defer to the PCF (using random number probabilities).
   4) If record from "New SSN List" and reflects a blank value, output a blank value to the CPR; if ties exist among the records, output "female" gender.

e. **Select Date of Death** (DOD) based on the following rules:
   1) If Medicare record reflects DOD, output the value.
   2) If more than one Medicare record reflects DOD, select the value from the most recent record (based on transaction cycle date).
   3) If no Medicare record available, output the value present on the Numident.
   4) If no reported DOD, defer to the PCF using random number probability after calculating gender.
   5) If input is blank and the PCF indicates "alive," output a blank DOD value.

f. **Select** the **date of birth** (DOB) based on the following rules:
   1) Select the highest DOB score within the following source file priority:
      a) Medicare
      b) Selective Service
      c) Census Numident
      d) HUD TRACS
      e) Indian Health Service
   2) If input is blank, output a blank value to the CPR.

g. **Select** the **best "name** fields" based on the following criteria:
   1) Highest name score with an exact match of last name.
   2) Exclude all IRS records and records from the "New SSN List."
   3) If only excluded names are in the processing array, select the first record read-in.
   4) If ties occur, select the first record read-in.

7. Each variable is flagged to reflect the decision rule invoked and the source of the data. Decision rules are established to account for the characteristics of each input source date.

Attachment 3. Description of FAV Status Codes

The FAV status code, a 5-character field, displays the ultimate resolution for all addresses based on findings from the FAV operation.

FAV Status Code. Parenthetical valid/invalid designation following the codes refer to how the address will be categorized in the application of an estimation formula to be applied in a later AREX address file.

Character 1 = Address Search Status (Found /Not Found/Unresolved)
0 = found (valid address)
1 = not in test site (invalid address)
2 = can't find/doesn't exist (invalid address)
3 = unresolved (invalid address)
4 = unresolved – junk (invalid address)
5 = unresolved – no other data (invalid address)
6 = unresolved – found basic street address but cannot confirm exact unit in a multi-unit structure (invalid address)
X = address not in FAV sample

Character 2 = Residential Status
0 = residential only (valid address)
1 = commercial only (invalid address)
2 = mixed residential/commercial (valid address)
3 = special place/group quarters (valid address)
4 = commercial mailbox service (invalid address)
N = not applicable
X = address not in FAV sample

Character 3 = Block-County Status (suffix changes are not addressed in this category)
0 = no change to block/county (valid address)
1 = block change (valid address)
2 = county change (invalid address)
3 = block and county change (invalid address)
N = not applicable
X = address not in FAV sample

Character 4 = Address Correction Status
0 = no change (valid address)
1 = basic street address (BSA) change – no unit designator (valid address)
2 = unit designator change, no change to BSA (valid address)
3 = BSA change, no change to unit designator (valid address)
4 = BSA change and unit designator change (valid address)
5 = listed as multi-unit, actually a single unit (valid address)
6 = listed as single unit, actually a multi-unit (invalid address)
N = not applicable
X = address not in FAV sample

Character 5 = Duplicate Status
0 = not a duplicate (valid address) (valid address)
1 = preferred duplicate (valid address)
2 = non preferred duplicate (invalid address)
N= Not applicable
X = address not in FAV sample